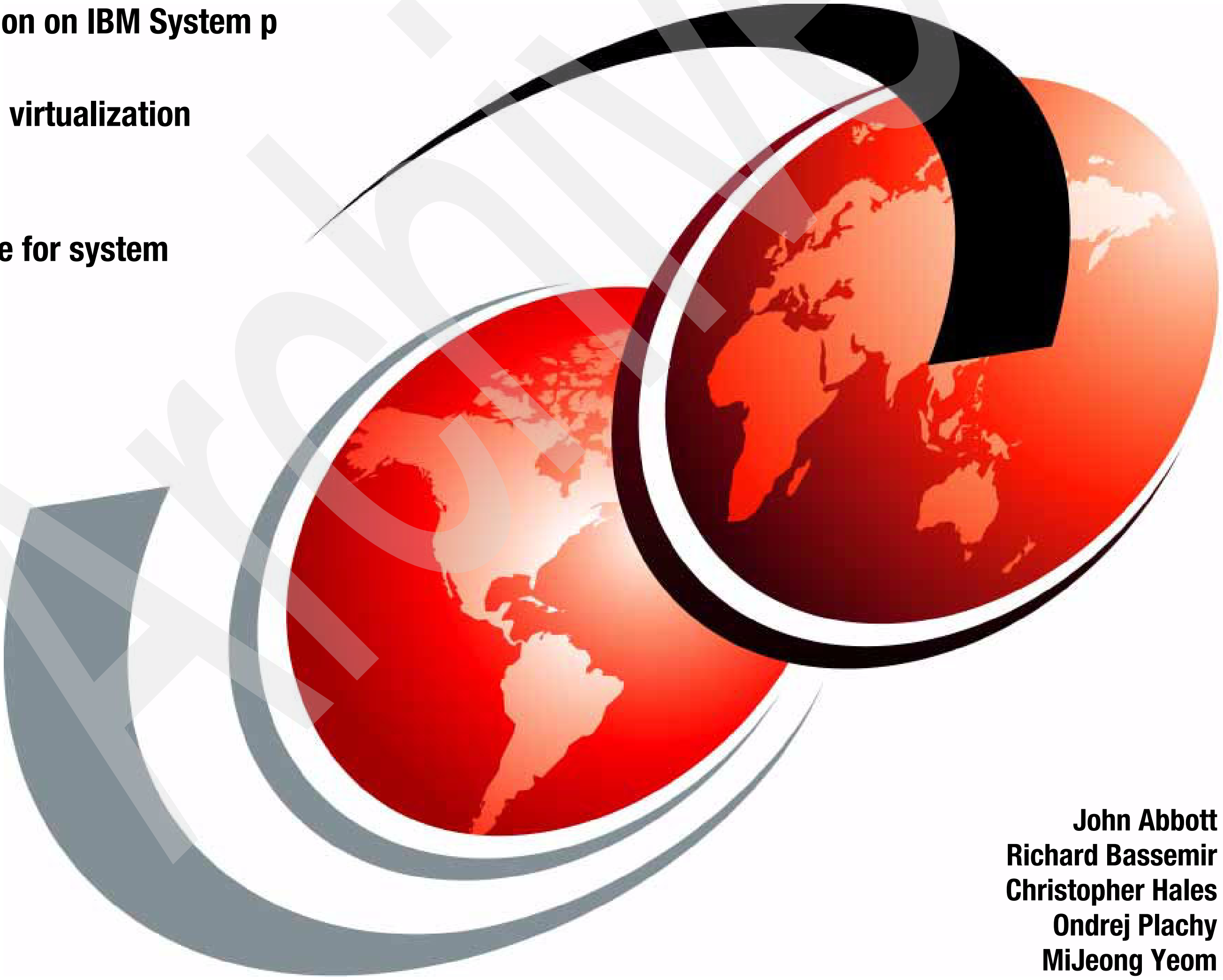
Advanced POWER Virtualization on IBM System p

Virtual I/O Server Deployment Examples

Scenarios for quick deployment of virtualization on IBM System p

Real world virtualization examples

A reference for system architects



John Abbott
Richard Bassemir
Christopher Hales
Ondrej Plachy
MiJeong Yeom

Redpaper

ibm.com/redbooks

International Technical Support Organization

Advanced POWER Virtualization Deployment Examples

February 2007

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (February 2007)

This edition applies to:
Version 1, Release 3, of the IBM Virtual I/O Server (product number 5765-G34)
Version 5, Release 3, technology level 5, of IBM AIX 5L for POWER (product number 5765-G03)
Version 240, Release 219, Modification 201, of the IBM POWER5 system firmware
Version 5, Release 2, Modification 1, with specific fixes MH00688 and MH00695, of the Hardware Management Console.

Contents

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX 5L™
AIX®
BladeCenter®
DS6000™
DS8000™
eServer™
HACMP™
IBM®
Micro-Partitioning™
POWER Hypervisor™
POWER5+™
POWER5™
POWER™
pSeries®
Redbooks (logo) ™
Redbooks™
System p5™
System p™
TotalStorage®
WebSphere®

The following terms are trademarks of other companies:

SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates.

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Preface

The *Advanced POWER Virtualization on IBM System p Virtual I/O Server Deployment Examples*, REDP-4224, publication provides a number of high-level system architecture designs using the Advanced POWER™ Virtualization feature available on IBM® System p5™ servers. These high-level architecture designs are referred to as *scenarios* and they show different configurations of the Virtual I/O Server and client partitions to meet the needs of various solutions.

The Advanced POWER Virtualization feature is very flexible and can support several configurations designed to provide cost savings and improved infrastructure agility. We selected the scenarios described in this paper to provide some specific examples to help you decide on your particular implementation and perhaps extend and combine the scenarios to meet additional requirements.

This publication is targeted at architects who are interested in leveraging IBM System p™ virtualization using the Virtual I/O Server to improve your IT solutions. System architects can use the scenarios as a basis for developing their unique virtualization deployments. Business Partners can review these scenarios to help them understand the robustness of this virtualization technology and how to integrate it with their solutions. Clients can use the scenarios and examples to help them plan improvements to their IT operations.

The team that wrote this Redpaper

This Redpaper was produced by a team of specialists from around the world working at the International Technical Support Organization, Austin Center.

John Abbott is a Senior Technical Analyst in Sydney, Australia. He has 10 years of experience in the UNIX® and network support field and has a diploma in IT. John has been team leader for UNIX Support in a leading Australian bank for five years and has been a driver on many strategic innovations using IBM System p technology.

Richard Bassemir is a Senior Software Engineer in the ISV Business Strategy and Enablement organization within the Systems and Technology Group in Austin, Texas. He has five years of experience in IBM System p technology. He has worked at IBM for 29 years. He started in mainframe design, design verification, and test, and moved to Austin to work in the Software Group on various integration and system test assignments before returning to the Systems and Technology Group to work with ISVs to enable and test their applications on System p hardware.

Christopher Hales is a Systems Architect and Certified IT Specialist based in the United Kingdom currently working in the Major Contracts Team of the IBM Systems and Technology Group. Chris has been designing and implementing IT solutions on behalf of clients for more than 20 years. He assists several large IBM clients to deploy Advanced POWER Virtualization and other virtualization technologies within their data centers. Chris holds an honors degree in computer science.

Ondrej Plachy is an IT Specialist in IBM Czech Republic responsible for project design, implementation, and support of large-scale computer systems. He has 12 years of experience in the UNIX field. He holds the Ing. academic degree in computer science from Czech Technical University (CVUT), Prague. He has worked at Supercomputing Centre of Czech Technical University for four years and has worked for IBM for eight years in the IBM AIX® 5L™ support team.

MiJeong Yeom is an IT Specialist in the Advanced Technical Support team of the Systems and Technology Group in Korea. She started at SI Company and moved into IBM to support ISVs to enable and test their solutions on System p technology. She is a Certified Advanced Technical Expert for System p and AIX 5L.

The project that created this publication was managed by:

Scott Vetter, PMP
IBM Austin

Thanks to the following people for their contributions to this project:

Ray Anderson, John Banchy, Lupe Brown, Erin Burke, Arzu Gucer, Bob Kovacs,
Jay Kruemcke, Bob Minns, Jorge Nogueras, Bret Olszewski, Maneesh Sharma,
Vasu Vallabhaneni
IBM U.S.

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and client satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this Redpaper or other Redbooks™ in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  ibm.com/redbooks

- Send your comments in an e-mail to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

1

Introduction

The *Advanced POWER Virtualization on IBM System p Virtual I/O Server Deployment Examples*, REDP-4224, publication provides a number of high-level system architecture designs using the Advanced POWER Virtualization feature available on IBM System p5 servers. These high-level architecture designs are referred to as *scenarios* and they show different configurations of the Virtual I/O Server and client partitions to meet the needs of various solutions.

The Advanced POWER Virtualization feature is very flexible and can support several variations designed to deliver cost savings and improved infrastructure agility. The scenarios covered in this paper are not the only ways to use this feature but they do give some specific examples to help decide on a particular implementation.

We discuss these scenarios in Chapter 2, “Virtual I/O Server scenarios” on page 13 in two sections.

- Section 2.1, “Virtual I/O networking” on page 14 concerns networking and shows different designs of the network using the virtual Ethernet features of the Virtual I/O Server.
- Section 2.2, “Virtual I/O Server SCSI” on page 45 focuses on virtual SCSI services and shows different designs of the disk storage using the virtual disk features of the Virtual I/O Server.

The different scenarios within each group vary depending on the degree of complexity. The complexity of the scenario is primarily determined by the a number of *architectural attributes*. The architectural attributes used by this paper are *resilience*, *serviceability*, *scalability/throughput*, and *flexibility*.

There are many combinations of configurations, and it would be difficult to document them all. The approach here is to provide some configurations in the form of these scenarios to be used as a starting point. For a complete virtualization solution, it is possible to combine scenarios that use both network and storage designs for a complete network and storage implementation.

Chapter 3, “Deployment case studies” on page 67 provides examples of how clients have used virtualization to solve real business problems and save costs.

1.1 Reader prerequisites

To effectively use this document, a basic understanding of the following areas as a prerequisite is helpful:

- Virtual I/O Server
- Logical partitions (such as concepts, creation, and management)
- Networking and disk storage concepts

For additional background information, refer to the following publications:

- For general Advanced POWER Virtualization information:
  - *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768
  - *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- For logical partitioning and dynamic logical partitioning:
  - *Partitioning Implementations for IBM eServer p5 Servers*, SG24-7039
- For setting up a basic Virtual I/O Server:
  - *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- For information about Integrated Virtualization Manager:
  - *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- For storage subsystem information:
  - *IBM TotalStorage DS300 and DS400 Best Practices Guide*, SG24-7121
  - *The IBM TotalStorage DS6000 Series: Concepts and Architecture*, SG24-6471
  - *The IBM TotalStorage DS8000 Series: Concepts and Architecture*, SG24-6452
- For information about the HMC:
  - *Effective System Management Using the IBM Hardware Management Console for pSeries*, SG24-7038

1.2 How to use this document

Although you can read this publication from start to finish, it is structured to facilitate quick reference and access to the most appropriate sections that address specific design requirements. Use the scenarios as a starting point and modify them to fit your specific implementation requirements.

The section for each scenario has a diagram, a high-level description, and an architectural attributes matrix to illustrate a possible configuration of the Virtual I/O Server. Each design has different architectural attributes that meet differing requirements. The purpose of the matrix is to communicate the authors insight as to the suitability of any particular scenario to a workload or deployment pattern being considered by a client.

Important: We strongly recommend that you always validate a scenario with actual workloads in a test environment before putting it into production.

1.2.1 Architectural attributes

We use four architectural attributes to describe the different scenarios. They are:

- Resilience

  Resilience is an indication of the ability to sustain operations in the event of an unplanned outage. For example, high resilience indicates that client partitions remain available in the event of a port or adapter failure.

- Serviceability

  Serviceability is an indication of how well the design can handle a planned maintenance of the Virtual I/O Server. High serviceability indicates that virtual networking and virtual SCSI services can remain available to client partitions while a Virtual I/O Server is maintained. Concurrent maintenance is possible with highly serviceable solutions.

- Scalability/throughput

  Scalability/throughput is an indication of how easily the design can scale and provide throughput. For example, a highly scalable configuration will easily allow the addition of new client partitions while containing the increased demand on the shared virtual network or shared virtual SCSI components.

- Flexibility

  Flexibility is an indication of how easy it is to add a client partition to the virtual infrastructure and participate in virtual services. It is an indication of the amount of unique (specific) systems-administration effort that would be required to add the new client partition or change an existing configuration.

Matrix rating notation

We give each architectural attribute in the matrix a rating to illustrate the robustness of the design. We use the following four levels of ratings:

- Standard

  Standard is the basic or default level. This rating illustrates that the design is using the inherent reliability, availability, and serviceability that comes with IBM System p5 technology.

- Medium

  Medium level indicates the scenario has some design features that are an improvement over standard and a range of associated benefits.

- High

  High level indicates that the architectural attribute is fully realized and exploited within the design. For example, dual Virtual I/O Servers enable excellent serviceability because one can be maintained while the other satisfies the workload requirements.

- Clustered

  We use a clustered rating if the architectural attribute is realized by exploiting IBM clustering technology that uses at least one other physical System p5 server in addition to the Advanced POWER Virtualization features, for example, clustered systems providing failover within the same data center. Clustering technology often provides exceptional levels of architectural protection.

1.2.2 Combining scenarios

It would be difficult to document all the different variations in design that might be considered using the Advanced POWER Virtualization features on the IBM System p5 server. The approach in this publication is to provide some basic scenarios for system architects to use as an initial starting point when considering the specific deployment for a project or client.

It is best to select a scenario that satisfies the network requirements and a scenario that satisfies the disk requirements. Then, combine the selected scenarios to arrive at a design for a network and storage Virtual I/O Server deployment.

To illustrate this concept, Chapter 3, “Deployment case studies” on page 67 is dedicated to examples about how these basic scenarios are combined to solve a workload needs.

1.2.3 Summary table of architectural attributes and ratings for all scenarios

The information in Table 1-1 on page 4 summarizes the architectural attributes and the corresponding ratings for each scenario. This enables the rapid identification of appropriate scenarios for particular deployment circumstances.

Table 1-1 Architectural attributes and ratings summary table

	Scenario	Title	Resilience	Serviceability	Scalability/ throughput	Flexibility	Reference
Virtual Ethernet Networking	1	One VIOS[a] with multiple LAN segments	Std	Std	Std	High	Section 2.1.2
	2	One VIOS with VLAN tagging	Std	Std	Std	High	Section 2.1.3
	3	One VIOS with client-partition backup link feature	Med	High	Std	Med	Section 2.1.4
	4	One VIOS with link aggregation and backup feature (type 1)	Med	Std	Std	High	Section 2.1.5
	4a	One VIOS with link aggregation and backup feature (type 2)	Med	Std	Med	High	Section 2.1.5
	5	Dual VIOS with client-partition backup link feature	High	High	High	Med	Section 2.1.6
	6	Dual VIOS with Shared Ethernet Adapter failover	High	High	High	High	Section 2.1.7
Virtual SCSI	7	One VIOS with single path to storage	Std	Std	Std	High	Section 2.2.2
	8	One VIOS with MPIO[b] access to storage	Med	Std	High	High	Section 2.2.3
	9	Dual VIOS with AIX 5L client mirroring (type 1)	High	High	Std	High	Section 2.2.4
	9a	Dual VIOS with AIX 5L client mirroring (type 2)	High	High	High	Med	Section 2.2.4
	10	Dual VIOS with MPIO and client-partition with MPIO	High	High	High	High	Section 2.2.5
	11	Dual VIOS with MPIO, client-partition with MPIO, and HACMP™	Clust	Clust	High	High	Section 2.2.6

[a] Virtual I/O Server

[b] Multipath I/O

1.2.4 Overview of common server workloads

There are various server workloads in the client IT environment that place different requirements on the underlying IT infrastructure. To satisfy these workloads, some infrastructures might require high network bandwidth, a high degree of backup, or have large disk I/O rates. Most workloads have a combination of these various characteristics.

This section describes the characteristics of some common workloads that might be useful when trying to categorize the deployment you are considering. When reviewing the scenarios in Chapter 2, “Virtual I/O Server scenarios” on page 13, it might be helpful to reflect on the characteristics of these workloads when making your own assessment.

Web servers

Web servers require high and reliable network bandwidth but, typically, do not require high disk I/O throughput. Also, they are often implemented in load-balanced clusters with automatic routing to systems within the cluster. There might be a case when an individual production Web server does not require high resilience. Often, there is a demand for the whole solution with Web servers to be horizontally scalable and flexible so that additional Web servers can be added as application load increases.

Application servers

Application servers usually require high network throughput and resilience, but their disk capacity and throughput requirements can vary substantially from low to very high. Usually, in production environments, they require high resilience and serviceability. In addition, they can also be implemented in *application clusters*, where the cluster is resilient to failure.

Database servers

Database servers usually require medium to high network throughput with resilience and very high disk throughput with additional reliability. In production environments, they are typically implemented in clusters. They require very high resilience and serviceability. In many implementations, some method of data replication for redundancy is implemented and data is often placed on external SAN devices.

Transaction middleware servers

Transaction middleware servers usually require very high network throughput and medium to high disk throughput. When implemented, they became an essential part of the IT infrastructure with very high availability requirements. Usually, in production environments, they are implemented in clusters.

File servers

File servers usually require high network throughput and high disk throughput. They can also be implemented in clusters in particular production environments.

Development and test servers

Servers that support development and test environments typically do not have large network throughput requirements and often do not require large disk throughput. These servers can be much more dynamic and thus can have a short life span. The ability to retask these servers brings financial and productivity benefits. In most cases, these servers do not require production level resilience and serviceability.

1.3 HMC and the Integrated Virtualization Manager

Using the Integrated Virtualization Manager facilitates the management of entry-level IBM System p5 servers through a Web browser interface. When using the Integrated Virtualization Manager, the Virtual I/O Server owns all of the adapters in the system. All client partitions use virtual network and virtual disk storage connections, which are serviced by the Virtual I/O Server. The scenarios in the following chapter that are targeted for entry level systems, such as the IBM System p5 505 or p5-510, might be suitable for use of the Integrated Virtualization Manager. If business needs require client partitions to have dedicated adapters or dual Virtual I/O Servers, the Hardware Management Console (HMC) is required.

> **Note:** The Integrated Virtualization Manager owns all of the network and disk resources. It allows each client partition to have a maximum of two virtual Ethernet connections and one virtual SCSI connection. If the client partition requires a dedicated adapter (network or disk), use the Hardware Management Console.

Using the Hardware Management Console allows greater flexibility with the scenarios in the following chapter. For example, with the Hardware Management Console, you are able to expand the scenarios by providing more than two virtual Ethernet connections to each client partition, adding additional Virtual I/O Servers to the system, or by being able to dedicate an adapter to a client partition.

1.4 High availability and networking setup of HMC

Some of the scenarios we present in Chapter 2, "Virtual I/O Server scenarios" on page 13 are constructed with high resilience and serviceability in mind. Most clients who manage several System p5 servers implement redundant HMCs to achieve these overall data center-wide service objectives. This section outlines some specific setups for the redundant HMCs and related networking considerations.

1.4.1 Redundant Hardware Management Consoles

Implementing redundant HMCs allows the continuation of important management functions to be performed during the maintenance cycle of one HMC. The HMC provides the following important management functions:

- Creating and maintaining a multiple partition environment
- Displaying a virtual operating system session terminal for each partition
- Displaying a virtual operator panel of contents for each partition
- Detecting, reporting, and storing changes in hardware conditions
- Turning on/off the managed partitions and systems
- Acting as a service focal point

While one HMC is not available, the other performs the management functions.

In addition, in an HACMP Version 5.3 high availability cluster software environment, HACMP communicates with the HMC to activate Capacity Update on Demand (CUoD) resources automatically (where available) during a take-over sequence. Therefore, the HMC is an integral part of the cluster. To ensure that the potential availability of the cluster is not degraded during HMC maintenance cycles, it is important to implement redundant HMCs.

1.4.2 Networking with redundant HMCs

IBM System p5 servers managed by HMCs require network connectivity between the HMC and the service processors of the System p5 servers. There can be redundant service processors in a mid- and high-end server (p5-570, p5-590, and p5-595) that require Dynamic Host Configuration Protocol (DHCP) and places more constraints on the HMC network architecture. Moreover, if dynamic partition operations are required, all AIX 5L and Linux® partitions must be enabled with a network connection to the HMC. For partitions in an IBM System p5 server, it is possible to use the Shared Ethernet Adapter in the Virtual I/O Server for a connection from the HMC to client partitions. Therefore, the client partition does not require its own physical adapter to communicate with an HMC.

As previously mentioned, some mid- and high-end IBM System p5 servers have two service processors. We recommend using both of the service processors for a redundant network configuration. Depending on the environment, there are several options to configure the network, and every deployment project will need to take the HMC networking requirements into consideration to properly plan for the network infrastructure.

We do not discuss all possible HMC networking configurations in this publication, but as an example of a redundant HMC network configuration, we present an overview design in Figure 1-1 on page 8. It provides one possible high availability configuration with two HMCs and two high-end IBM System p5 servers. Note that three distinct layer-2 Ethernet networks are needed to provide the required functionality. Two of them are private (closed) networks, built specifically for IBM System p5 hardware management. No other systems must be connected to the HMC LANs; they are managed by DHCP servers running on respective HMCs. The third network is a usual management network for management access to the HMCs and partition from the client network.

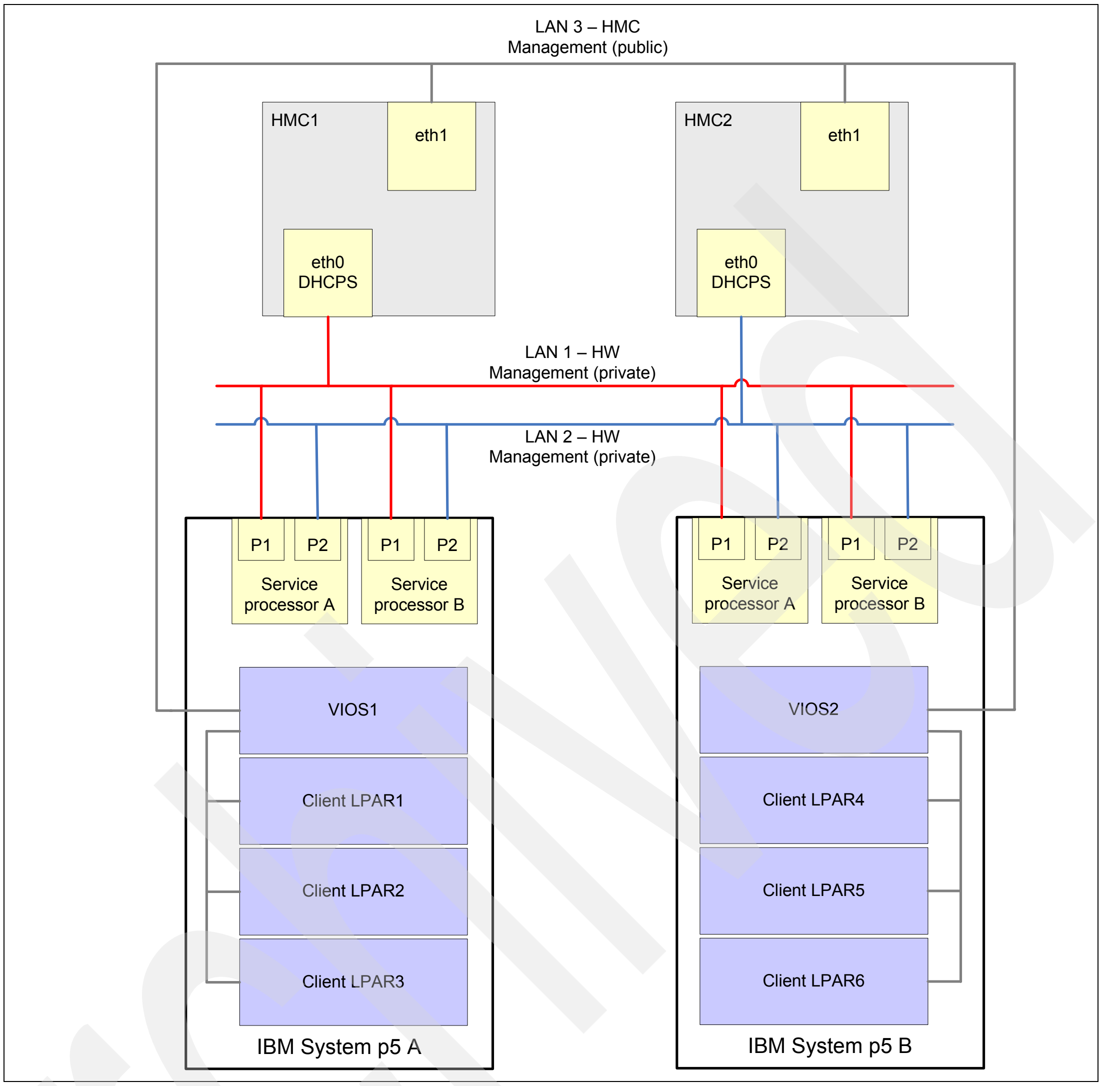


Figure 1-1 Highly available networking with two HMCs

1.5 Micro-Partitioning technology

With the introduction of the POWER5™ processor, partitioning technology moved from a dedicated processor allocation model to a virtualized, shared processor model. This section briefly discusses shared processor logical partitioning (IBM Micro-Partitioning™ technology) on IBM System p5 servers.

Micro-Partitioning technology has been available for a considerable period, is well understood in most client environments, and is well documented elsewhere. Consequently, this document neither describes the Micro-Partitioning technology in detail nor presents specific scenarios focused on Micro-Partitioning technology.

The Micro-Partitioning feature on System p5 servers provides the capability to create multiple partitions that share processor resources. The minimum processor entitlement for a micro-partition is 1/10th of a CPU (with 1/100 granularity once the minimum is met). Any of the virtual partitions can run on any of the physical processors in the shared pool.

Micro-partitions are either *capped* or *uncapped*. Capped partitions can never use more CPU cycles than their entitlement (but can cede or donate unused processor cycles back to the POWER Hypervisor™ for reallocation). Uncapped partitions, however, can have additional processor cycles donated to them by the POWER Hypervisor if they demand CPU and the processor pool has the spare resources to allocate. The weighting of the uncapped micro-partitions that are competing for additional processor cycles determines the actual number of additional cycles the POWER Hypervisor will allocate to them above their entitlement.

This dynamic allocation of processor resources to partitions with runable workload demands makes it possible to run the physical server at very high utilization levels.

To achieve the virtualization of processor resources, physical processors are abstracted into virtual processors and made available to partitions. The POWER Hypervisor controls the dispatch of virtual processors to physical processors.

The main advantages of Micro-Partitioning technology are:

- More partitions

  The granularity of a dedicated, whole processor might be too coarse. Not every application requires the full processing power of a POWER5+™ processor. Using Micro-Partitioning technology, there can be up to 10 times more partitions than physical CPUs (up to a maximum of 254 partitions).

- Better utilization of physical CPUs

  Usually, with dedicated processors, partitions and stand-alone servers are planned to have around 40% spare processing power to cover workload peaks and increasing application demands over time. With a shared processor pool and micro-partitions, the workload peaks in individual partitions can be more efficiently served. The planned overall CPU utilization in a IBM System p5 server using Micro-Partitioning technology can usually be higher than that of stand-alone servers without affecting service levels.

Here we provide a basic example to demonstrate some advantages of Micro-Partitioning technology. Assume that you have to build a small system with one Web server, one application server, and one LDAP directory server on an IBM System p5 platform with four physical processors (Figure 1-2 on page 10). The sizing is:

- In dedicated mode (or stand-alone servers):

Web server:	1 CPU
Application server:	2 CPUs
LDAP server:	1 CPU

  This, potentially, can result in unused processor cycles in both the Web server and LDAP server and saturated CPU load in the application server.

- Using Micro-Partitioning technology:

Web server:	CPU entitlement 0.5, uncapped
Application server:	CPU entitlement 3 CPUs, uncapped
LDAP server:	0.5 CPU, uncapped

  Because all the partitions are uncapped, the actual processing power will be distributed by IBM System p5 POWER Hypervisor according to the applications' requirements. The actual CPU resource distribution will be dynamic and can result in these values, for example: Web server 0.6, application server 3.2, LDAP server 0.2.

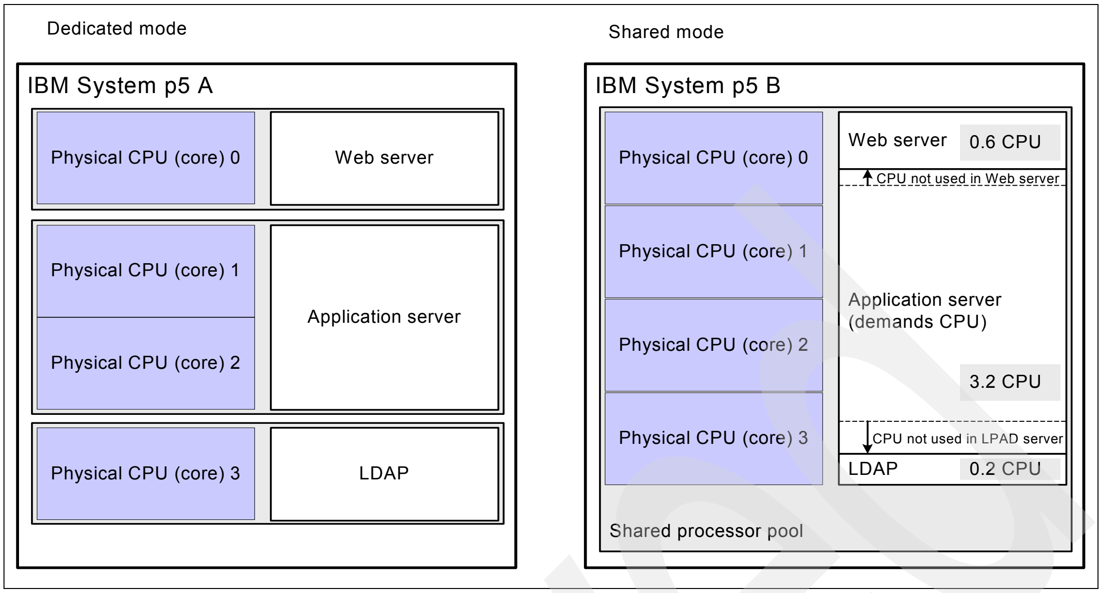


Figure 1-2 Dedicated and shared CPU comparison

Using Micro-Partitioning technology, the CPU usage is more efficient and applications can use the available processing resource when they need it. Therefore, the overall throughput of the whole example solution is higher.

1.5.1 Virtual I/O Server and Micro-Partitioning technology

Although dedicated processor partitions offer better memory affinity, the shared processor partitions have the advantage by dynamically growing and shrinking processor resources for particular partitions as the workload varies. In general, dedicated and shared processor partitions provide equivalent quality of service.

Appropriate CPU entitlement sizing is important for workloads with response time requirements (such as the Virtual I/O Server). A very low CPU entitlement for the Virtual I/O Server can cause longer Virtual I/O Server response times in situations when the shared processor pool is highly utilized. Therefore, such a situation must be avoided.

The processing requirements associated with a particular deployment for virtual SCSI and virtual Ethernet vary substantially. If both the virtual SCSI and SEA adapter (bridging between internal and external network) require very high throughput, it might be best to consider implementing separate Virtual I/O Server partitions for each type of service and tuning the processor resource allocations accordingly.

1.5.2 Considerations for Virtual I/O Server resources

When deploying Virtual I/O Server in a shared processor partition, we recommend:

- A good understanding of the CPU requirements for each workload type (virtual network and virtual SCSI) and allocating the appropriate CPU entitlement. Under most circumstances, close to the minimum entitlement will not be the best starting point (remember that if not all the Virtual I/O Server partition entitlement is used, the spare cycles can be reallocated to other partitions; therefore, it is safe to allocate more CPU entitlement to the Virtual I/O Server than actually needed).
- Configuring the Virtual I/O Server partition as uncapped with a high weight so that it can receive additional CPU cycles if handling larger workloads (if spare cycles are available in the processor pool).
- Tuning the networking. Use *large send* if possible, allow MTU discovery, and allow threading mode if virtual SCSI is used in the same Virtual I/O Server.
- Thoroughly testing the performance under load before placing it into production.

For further information regarding Virtual I/O Server resource allocation and setup, refer to *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194.

2

Virtual I/O Server scenarios

In this chapter, the basics explained in Chapter 1 are applied to Virtual I/O Server scenarios, which, although useful in their current form, are designed to form the basic building blocks of complex systems that can be applied to production environments.

We divide these scenarios into two main categories:

- Section 2.1 “Virtual I/O networking” on page 14
- Section 2.2 “Virtual I/O Server SCSI” on page 45

2.1 Virtual I/O networking

This section describes possible virtual network scenarios. These scenarios are just a sample of the ways to configure the Virtual I/O Server to provide network resources to the client partitions.

2.1.1 Introduction to virtual I/O networking terminology

The POWER Hypervisor implements a VLAN-aware Ethernet switch that can be used by client partitions and the Virtual I/O Server to provide network access. This section outlines the basic terminology used in both the text and figures.

Figure 2-1 introduces the Virtual I/O networking terminology as it refers to the client partitions, the Virtual I/O Server, and other elements such as Ethernet switches and the POWER Hypervisor.

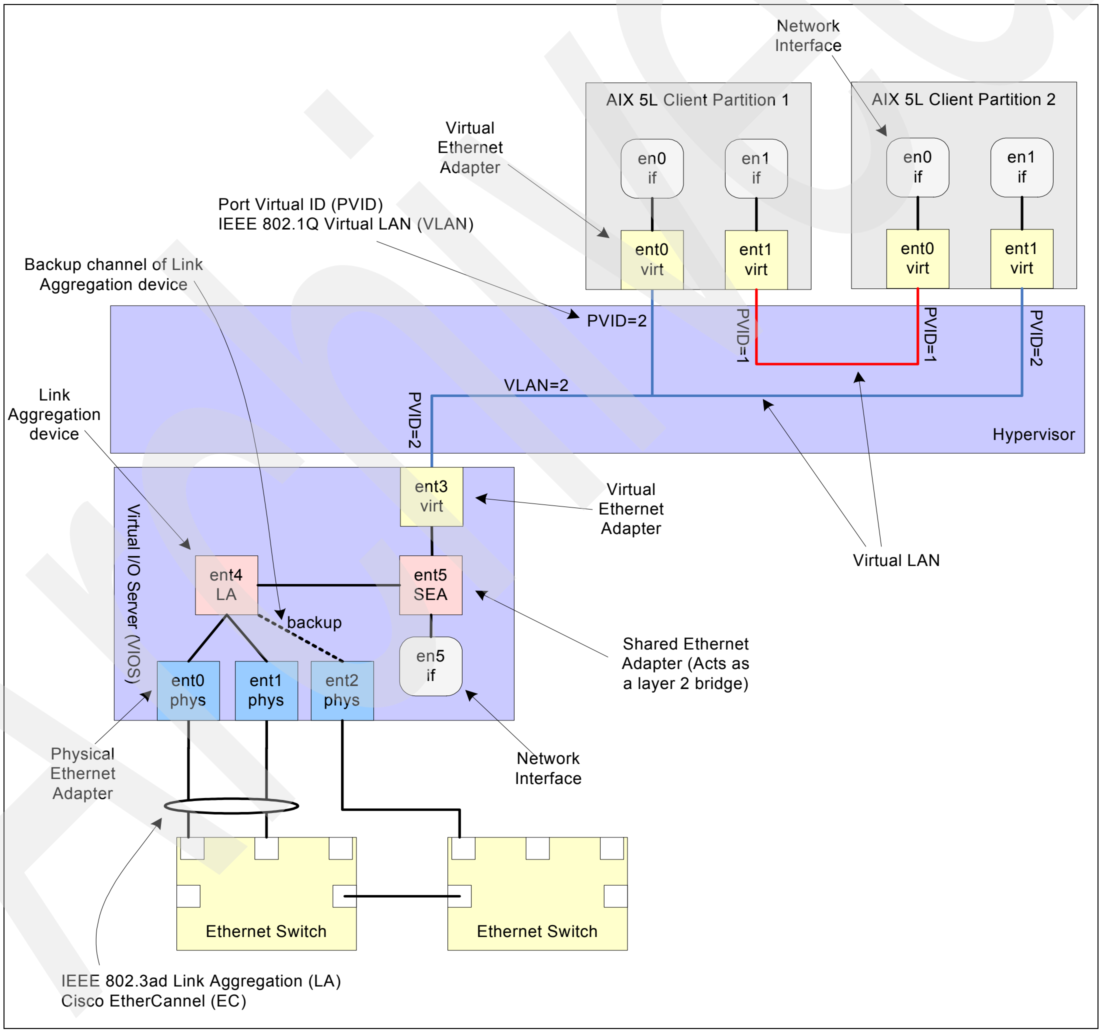


Figure 2-1 Virtual I/O network terminology

Client partition terminology

For most of the scenarios, the client partitions have few networking components. Each Ethernet adapter (such as device *ent0 virt*) is associated with one network interface (such as *en0 if*). An IP address can be defined on a network interface.

In a subset of the scenarios, the client partition network interface has two virtual adapters associated with it. One network adapter is the primary network connection, and the second is a backup connection. The backup adapter connection is shown in diagrams as a dashed line that is implemented in association with the link aggregation device.

Figure 2-1 on page 14 uses the following components and terms:

en0 if Ethernet interface providing programmed access to the virtual Ethernet adapter

ent0 virt Virtual Ethernet adapter that connects to the POWER Hypervisor VLAN-aware Ethernet switch

Virtual I/O Server terminology

There are several components that can be used to construct the network services of the Virtual I/O Server, depending on type of services required. Figure 2-1 on page 14 uses the following components and terms:

en5 if Ethernet interface providing programmed access to a Shared Ethernet Adapter.

ent3 virt Virtual Ethernet adapter that connects the Virtual I/O Server with the relevant VLAN used in the POWER Hypervisor switch.

ent5 SEA Shared Ethernet Adapter (SEA) device. This device can connect directly to a physical Ethernet adapter and functions as a layer-2 bridge transferring packets from the virtual Ethernet adapter to the physical Ethernet adapter. It can also connect and transfer traffic to a link aggregation device.

ent4 LA Link aggregation device. This is a component that implements IEEE 802.3ad Link Aggregation or Cisco EtherChannel. This link aggregation device will typically have two (maximum of eight) physical Ethernet adapters connected and will manage them as an single Ethernet connection.

ent0 phys Dedicated physical Ethernet adapter that connects to the physical network.

Backup The backup adapter of the link aggregation device is a connection between the interface and a second Ethernet adapter that provides a secondary path to the network if the primary path has a network failure.

Additional terms

Additional terms used in the scenarios include:

VLAN and VLAN ID Virtual LAN and VLAN identifier. VLAN technology establishes virtual network segments over Ethernet switches (such as the switch implemented by the POWER Hypervisor). A VLAN is normally a broadcast domain where all partitions and servers can communicate without any level-3 routing or inter-VLAN bridging and is denoted by a VLAN ID (for example, VLAN 100). VLAN support in AIX 5L V5.3 is based on the IEEE 802.1Q VLAN implementation. For a more information about VLAN technology, see 2.1.3 “One Virtual I/O Server with VLAN tagging: Scenario 2” on page 21.

PVID The default VLAN ID is also known as the port virtual ID (PVID) and identifies a port on a VLAN-capable switch (such as those implemented in the POWER Hypervisor) that indicates the default VLAN ID of the port.

Tagged and untagged IEEE 802.1Q defines two types of VLAN traffic: tagged and untagged packets. Tagged packets have a VLAN tag added to the Ethernet frame by the switch (if it is not already tagged), while untagged packets do not. The Ethernet switch makes switching decisions based on the contents of these tags. For untagged traffic, the default VLAN ID determines the default VLAN ID for which the packet is destined.

Overview of the link aggregation technology and features

Link aggregation is a network port aggregation technology that allows several network adapters to be aggregated together and behave as a single pseudo-network adapter. All Ethernet adapters are treated as a group and have one hardware address and one IP address. For this technology, all physical Ethernet adapters that are part of the same link aggregation must be connected to the same network switch. This switch might need to be configured for link aggregation unless they support the Link Aggregation Control Protocol (LACP). If one Ethernet adapter of the group fails, the other available Ethernet adapters can keep network communication active.

There are two types of link aggregation supported:

- Cisco EtherChannel

  EtherChannel is a Cisco-specific developed technology that requires specific network switch configuration.

- IEEE 802.3ad Link Aggregation

  IEEE 802.3ad is an open standard that might not need a network switch configuration if it implements LACP.

It is possible to use link aggregation technology to implement a backup network connection if the primary connection fails. This technology is named link aggregation with a backup channel. Under normal conditions, all network traffic passes through the primary adapters, and no traffic passes through the backup channel and adapter. The backup becomes active, with traffic flowing through it, only if the primary connection fails.

No additional configuration is required on a network switch to implement link aggregation with a backup channel. This technology allows greater network redundancy and provides higher network availability.

In general, link aggregation technology is used for:

- Greater reliability for network connectivity

  Multiple physical Ethernet adapters consist of one pseudo-network adapter. With this configuration, there is no disruption of the network.

- Greater network bandwidth
- Greater network resilience using the backup channel facility

With link aggregation, network traffic can be distributed between several network adapters to reduce the network throughput limitation of a single network adapter. This can be implemented to avoid a network bottleneck. However, when the link aggregation backup channel is used, the backup channel can only use one network adapter. Therefore, the link aggregated backup channel cannot deliver the same throughput as an aggregated connection made up of several network adapters that function as a single connection.

Further reference

For additional information, refer to the following sources:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768
- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *Partitioning Implementations for IBM eServer p5 Servers*, SG24-7039
- *IBM System p5 Approaches to 24x7 Availability Including AIX 5L*, SG24-7196

2.1.2 One Virtual I/O Server with multiple LAN networks: Scenario 1

Figure 2-2 depicts the network configuration for this scenario.

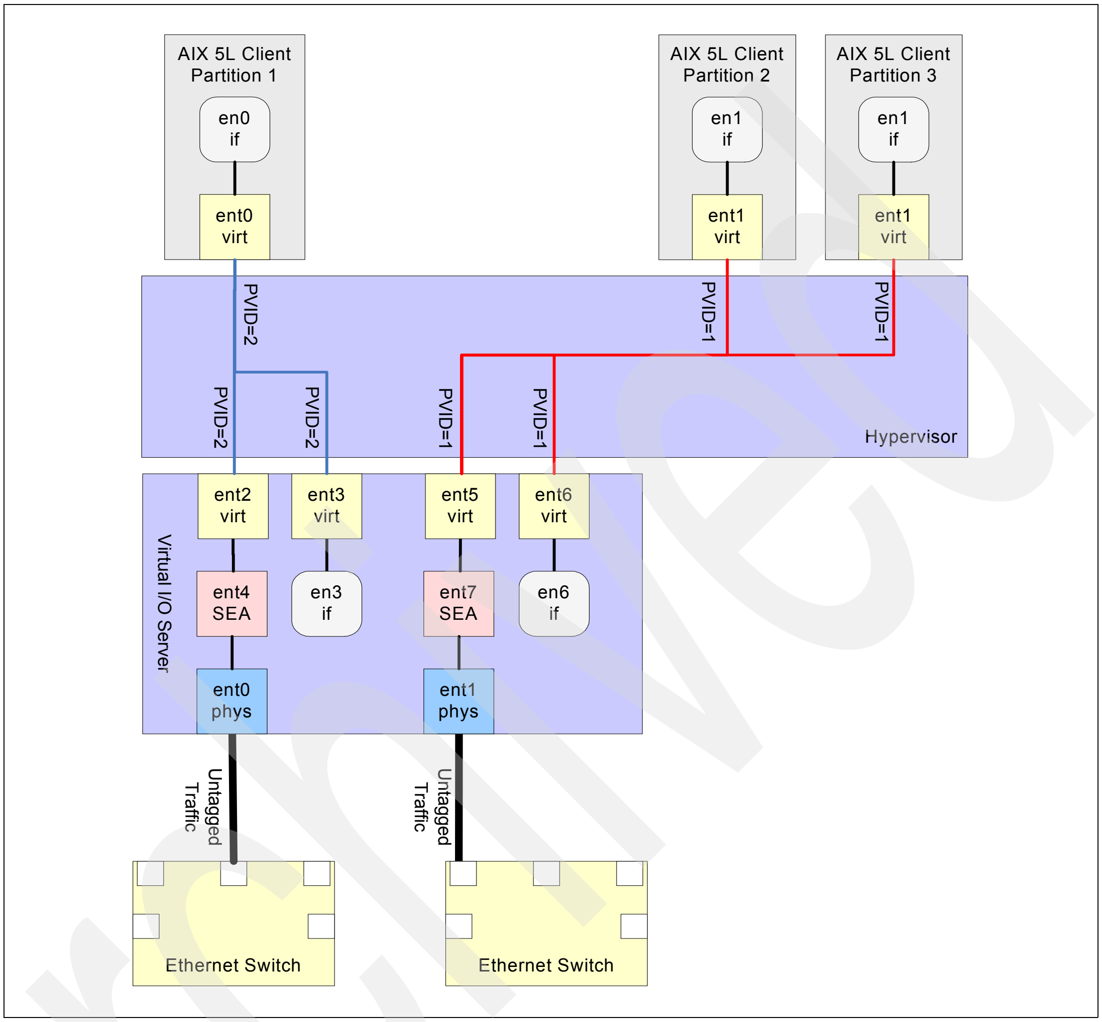


Figure 2-2 One Virtual I/O Server with two LAN networks

Description of scenario

In this configuration, a single Virtual I/O Server provides multiple network connections to the external network, one per VLAN, to the client partitions. There is no client requirement for a highly available network connection. This configuration is well-suited for entry-level systems that have dynamic requirements. For example, if brief code development, testing, or porting is required, this configuration lends itself to the quick setup and deployment of a client partition. For this basic configuration, either the Integrated Virtualization Manager or the Hardware Management Console (HMC) can be used.

In Figure 2-2, the Virtual I/O Server uses two physical Ethernet adapters and two Shared Ethernet Adapters to connect with two different VLANs to the external network. Using two different VLANs allows this configuration to support different network segments to the client partitions. For example, VLAN ID 1 can be on the 10.10.10.x network and VLAN ID 2 can be on the 10.10.20.x network. The POWER Hypervisor provides the Ethernet switch between the client virtual adapters and the Virtual I/O Server virtual adapters. With a single physical Ethernet adapter for each VLAN, an adapter, cabling reconfiguration, or hardware replacement will interrupt the client connection to the network.

If the Virtual I/O Server is taken offline for maintenance, the connection to the external network is unavailable and the client partitions will temporarily lose their network access during the maintenance. With a single physical adapter that is shared for each virtual Ethernet segment, an adapter failure will result in an outage for that virtual Ethernet network.

This scenario provides flexible network access for the client partitions. Refer to the other scenarios for designs with additional network resiliency.

Architectural attributes

Table 2-1 provides the architectural attributes.

Table 2-1 One Virtual I/O Server with two LAN networks

	Network access	
Resilience	Standard	🟩
Serviceability	Standard	🟩
Scalability/throughput	Standard	🟩
Flexibility	High	🟩🟩🟩

Note the following explanation of the attributes and their ranking:

- Resilience

  Standard
  This scenario provides the resilience inherent in each virtualization component.

- Serviceability

  Standard
  The use of one Virtual I/O Server does not provide the client partitions an alternative network connection when the Virtual I/O Server is offline for maintenance.

- Scalability/throughput

  Standard
  With this configuration, additional clients can be added to scale up the configuration without any having to impact the other partitions as long as the additional client partitions do not exceed the network adapter bandwidth. The throughput is determined by the single Ethernet adapter for each VLAN.

- Flexibility

  High
  This configuration provides the flexibility to add client partitions without additional required hardware. Both the HMC and the Integrated Virtualization Manager interfaces are tools to provision additional client partitions.

Summary

This scenario is well-suited for fundamental deployments of client partitions that need access to a network. It is an efficient way to make use of limited Ethernet adapters across multiple client partitions without a dedicated network connection. Adding and removing client partitions is straightforward. Multiple networks can be defined to satisfy the needs of client partitions connecting to different networks. The number of client partitions this scenario can support is determined by the aggregation of all client partition network traffic and the capability of the physical Ethernet adapters.

This scenario can be implemented with the Integrated Virtualization Manager or the Hardware Maintenance Console.

This configuration is suitable for solutions with the following characteristics:

- There is no requirement for a highly available network.
- The collective network traffic to the external network is moderate.
- There is a requirement for network separation between the client partitions, but there is no VLAN-aware network equipment.
- Dual Virtual I/O Servers are not desired.
- Entry-level systems that have dynamic requirements such as test and sandbox projects.
- Web, application, and database servers if there are no requirements for a high throughput and resilient network.

Solutions with the following characteristics are not recommended:

- When external network communication is critical.
- When most of the network traffic to the external network comes from only one client partition, and the network traffic from a client partition severely affects the total network performance.
- When most of the client partitions generate a large amount of network traffic simultaneously.

Further reference

For additional information, see the following references:

- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940

2.1.3 One Virtual I/O Server with VLAN tagging: Scenario 2

Figure 2-3 shows the Advanced POWER Virtualization architecture using the Virtual I/O Server for this scenario.

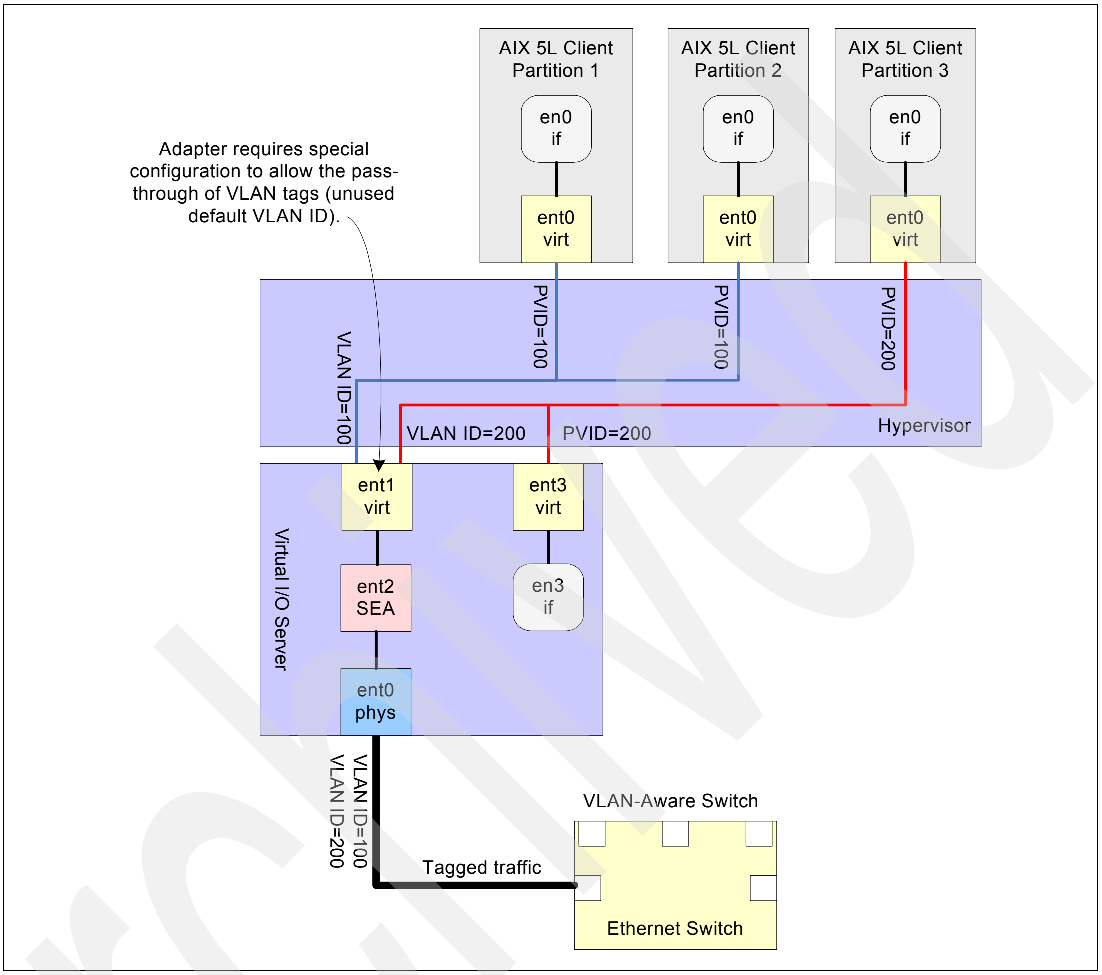


Figure 2-3 One Virtual I/O Server with multiple LAN segments (and VLAN tagging)

Description of the scenario

In this scenario, one Virtual I/O Server with a Shared Ethernet Adapter (SEA) enabled allows clients belonging to multiple VLANs to connect to an external network over one shared physical adapter hosted by the Virtual I/O Server. In order to fully describe this setup, it is helpful to define selected VLAN terminology and to describe how it is used with IBM System p5 Advanced POWER Virtualization feature.

VLAN technology overview

VLAN is an acronym for *virtual local area network*. VLAN processing is an extension of layer-2 network switching, and adds a four-byte *tag* with a *VLAN identifier* to each Ethernet frame (packet) transferred through a VLAN-aware switch. VLAN-aware switches scan for this tag and make switching decisions based on it (they interpret each VLAN transfer as a distinct layer-2 network; no packets are transferred between VLANs). The VLAN standard (IEEE-802.1Q) defines two types of traffic: untagged and tagged. Untagged packets are packets without a VLAN tag. VLAN-aware switches handle packets of both types to support VLAN-aware and non-VLAN aware devices.

To support devices without a VLAN ID, the VLAN specifications define a port-based assignment where each port is a member of a particular Virtual LAN (Port Virtual LAN ID also known as PVID). All untagged traffic going through a particular port belongs to the ports default Virtual LAN ID. Although most devices can be configured to accept and understand the tags, they are usually completely unaware of VLAN tagging at the switch and do not accept packets with VLAN tags.

The ports within the switch can belong to more than one VLAN with just one default VLAN ID for the port. Packets that arrive at the switch with a VLAN tag are allowed to pass through as long as the switch port is a member of the target VLAN. If a packet arrives at a switch port without a VLAN tag, the switch tags the packet with the default VLAN ID of the port. When a VLAN-aware switch sends a packet to a port with a VLAN tag matching the port's default VLAN ID, it strips the tag from the packet before sending it out of the switch to the attached device. If the VLAN tag matches one of the possible additional VLANs that the port belongs to (not the default VLAN ID), the tag will be left intact. This allows devices that are both VLAN aware and VLAN unaware to exist side-by-side.

Because the IBM System p5 POWER Hypervisor implements a VLAN-compatible switch with all the features described previously, partitions can use either the VLAN-aware approach or the VLAN-unaware approach. Therefore, for a VLAN-unaware device (such as AIX 5L and Linux in their standard configuration), you can connect to a VLAN by assigning the virtual switch port to (by virtual Ethernet adapter) the default VLAN ID that matches the VLAN ID. In simple scenarios, administrators do not have to configure VLAN tagging. In complicated scenarios, where connections to multiple external VLANs are required, VLAN-aware configurations might be needed.

Note: The IBM System p5 POWER Hypervisor virtual Ethernet is designed as a VLAN-aware Ethernet switch. Therefore, the use of virtual Ethernet in IBM System p5 servers uses VLAN technology. The default VLAN ID tag is added by the hypervisor to outgoing packets from a non-VLAN-aware partition. These tags are used to decide where the packet goes and are stripped off when the packet is delivered to the target partitions with the same default VLAN ID. Partitions do not have to be aware of this tagging or untagging.

To exchange packets between switches that participate in multiple VLANs, there is a need to transfer packets belonging to multiple VLANs over one shared channel. The term *trunk* is often used for such a network link carrying multiple VLANs. This link must be configured between tagged ports (IEEE-802.1Q) of VLAN-aware devices. Therefore, they are used in switch-to-switch links rather than links to hosts. In the virtual environment of IBM System p5 servers, a link to an external switch is necessary in order to be able to transfer traffic from multiple POWER Hypervisor-based VLANs to external VLANs over one physical interface. To implement the link, which is provided by the Virtual I/O Server's SEA feature, the virtual Ethernet adapter requires careful configuration in the Virtual I/O Server (access External Network and IEEE-802.1Q settings enabled), and the appropriate external switch port configuration is needed.

The Virtual I/O Server provides the connection between internal and external VLANs. There can be several VLANs within the same server (implemented by the POWER Hypervisor) that need to be connected to multiple VLANs outside the server. Therefore, the Virtual I/O Server must be connected to every VLAN. This does not allow packets to move between the VLANs.

The following features enable the multi-VLAN link between internal and external LANs:

- VLAN tags must not be stripped from packets.

  For example, IEEE-802.1Q protocol must be enabled on both the external switch port and the Virtual I/O Server's virtual adapter. VLAN IDs that communicate externally must not match default VLAN IDs of the Virtual I/O Server's virtual Ethernet adapter because the tags are automatically stripped off by the POWER Hypervisor. To avoid this, configure an unused VLAN ID on the Virtual I/O Server's virtual Ethernet adapter as the default VLAN ID and configure additional active VLAN IDs as additional LAN IDs on that adapter.

- The Shared Ethernet Adapter in the Virtual I/O Server (SEA) must be able to bridge packets for many VLANs.

  The SEA functions as an L2 bridge and transfers packets intact. Therefore, the VLAN tags get transferred between the physical and virtual sides.

To connect the Virtual I/O Server to the network, two approaches are possible:

- Defining an IP address on the SEA adapter
- Configuring an additional client virtual Ethernet adapter on the Virtual I/O Server

In most cases, defining a secondary virtual adapter and configuring the IP addresses is preferred instead of configuring an IP address on the SEA adapter. In multiple VLAN scenarios, this is a better approach, and it is more apparent to which VLAN the Virtual I/O Server is connected.

Description of the diagram

In Figure 2-3 on page 21, there are three client partitions and one Virtual I/O Server using a SEA on one physical adapter. The client partitions are connected to two VLAN networks (VLAN ID 100 and 200). These two VLANs are connected externally, and the VLAN tagging is managed by the POWER Hypervisor. The tagging is transferred intact to the external switch so that it can communicate with external devices in corresponding VLANs. The external switch port must allow tagged traffic.

The following points describe this scenario:

- Client partition setup

  The client partition has the following attributes:

  - AIX 5L client partition 1: Client virtual Ethernet adapter with default VLAN ID (PVID in earlier versions) 100
  - AIX 5L client partition 2: Client virtual Ethernet adapter with default VLAN ID 100
  - AIX 5L client partition 3: Client virtual Ethernet adapter with default VLAN ID 200

  With the previous configuration, partitions 1 and 2 are connected to VLAN 100 and partition 3 to VLAN 200. Partitions 1 and 2 can communicate directly each other. Partition 3 cannot communicate with either of the other two partitions.

- Virtual I/O Server setup

  The Virtual I/O Server has the following attributes:

  - Server virtual Ethernet adapter (ent1):
    - Default VLAN ID: An unused value, for example, 999
    - Additional VLAN IDs: 100, 200
    - IEEE-802.1Q enabled
    - Access to the external network (referred to as trunk previously) enabled

- External link: Physical Ethernet adapter (ent0): No special settings; IEEE-802.1Q (trunk port) must be enabled when connecting to the external switch port.
- SEA (ent2): Defined on ent0 and ent1 adapters.
- Client virtual Ethernet adapter (ent3): Default VLAN ID 100.

With these settings, tagged packets from both VLANs 100 and 200 are transferred by the SEA over the physical adapter and routed by the external switch to their respective VLANs. The Virtual I/O Server is connected by its virtual Ethernet adapter to VLAN 200 and can communicate directly with partition 3 and through the SEA with external devices on VLAN 200. It cannot communicate with partitions 1 and 2.

This scenario can be combined with the scenario described in 2.1.7 “Dual Virtual I/O Servers with SEA failover: Scenario 6” on page 40 to provide higher resiliency and VLAN tagging support.

Architectural attributes

Table 2-2 provides the architectural attributes.

Table 2-2 One VIOS with multiple LAN segments (and VLAN tagging) non-functional features

	Network access
Resilience	Standard 🟩
Serviceability	Standard 🟩
Scalability/throughput	Standard 🟩
Flexibility	High 🟩🟩🟩

Note the following explanation of the attributes and their ranking:

- Resilience

  Standard
  If the Virtual I/O Server is not available (for example, for maintenance), communication with the external network will not be possible. The architecture can be enhanced to increase resiliency by combining this scenario with the scenario described in 2.1.7 “Dual Virtual I/O Servers with SEA failover: Scenario 6” on page 40 and implementing two Virtual I/O Servers.

- Serviceability

  Standard
  If the Virtual I/O Server is taken offline for maintenance, communication from the client partitions to the external network will not be possible. Serviceability can be increased to high by combining this scenario with the scenario described in 2.1.7 “Dual Virtual I/O Servers with SEA failover: Scenario 6” on page 40.

- Scalability/throughput

  Standard
  Client partitions can be added up to the total aggregated throughput of a single Ethernet adapter. Traffic throughput can be increased further by combining this scenario with the scenario described in 2.1.5 “One Virtual I/O Server with link aggregation and backup link: Scenario 4” on page 29 and using link aggregation on the physical network adapters hosted by the Virtual I/O Server.

- Flexibility

  High
  Client partitions can be easily added without additional hardware (if performance throughput of single physical adapter is not overcommitted). Additional VLANs can be easily added.

Summary and applicable workloads

This scenario provides a flexible networking for complex networks. The POWER Hypervisor implements a VLAN-compliant switch and the SEA provides an inter-switch link (ISL) for communicating with multiple VLANs over one channel to the outside network. Apply appropriate planning of the Virtual I/O Server maintenance windows in production environments because external communication is not possible. It is important to consider that all traffic shares one physical link, and therefore, appropriate performance and capacity planning is required.

This scenario can be combined with 2.1.7 “Dual Virtual I/O Servers with SEA failover: Scenario 6” on page 40 for higher resiliency and with 2.1.5 “One Virtual I/O Server with link aggregation and backup link: Scenario 4” on page 29 for higher throughput.

This scenario can be used with many workloads where multiple VLANs are required. Solutions based on this scenario can range from low throughput to standard throughput (keep in mind that higher throughput is possible with link aggregation that uses multiple physical adapters).

This scenario is suitable for solutions with the following characteristics:

- When it is required to connect several VLANs between external and internal (virtual) network over the shared physical link.
- There is no requirement for a highly available network.
- Limited Ethernet adapters share multiple client partitions, and network traffic to the external network is moderate.
- There is a requirement for network separation between the client partitions and VLAN-aware network configuration is used.
- A test and development environment.
- A Web, application, and database server if there is no requirement of high throughput and resilience network.

Solutions with the following characteristics are not recommended:

- External network communication is critical.
- Most of the network traffic to external network comes from only one client partition (for example, a file server), and the network traffic from a client partition severely affects the total network performance.
- All the client partitions generate a large amount of network traffic to the external network simultaneously.

Further reference

For a more detailed discussion about VLAN technology and its implementation in IBM System p5 Advanced POWER Virtualization feature, refer to the following publications:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768

2.1.4 One Virtual I/O Server with client-partition backup link: Scenario 3

Figure 2-4 depicts the network configuration for this scenario.

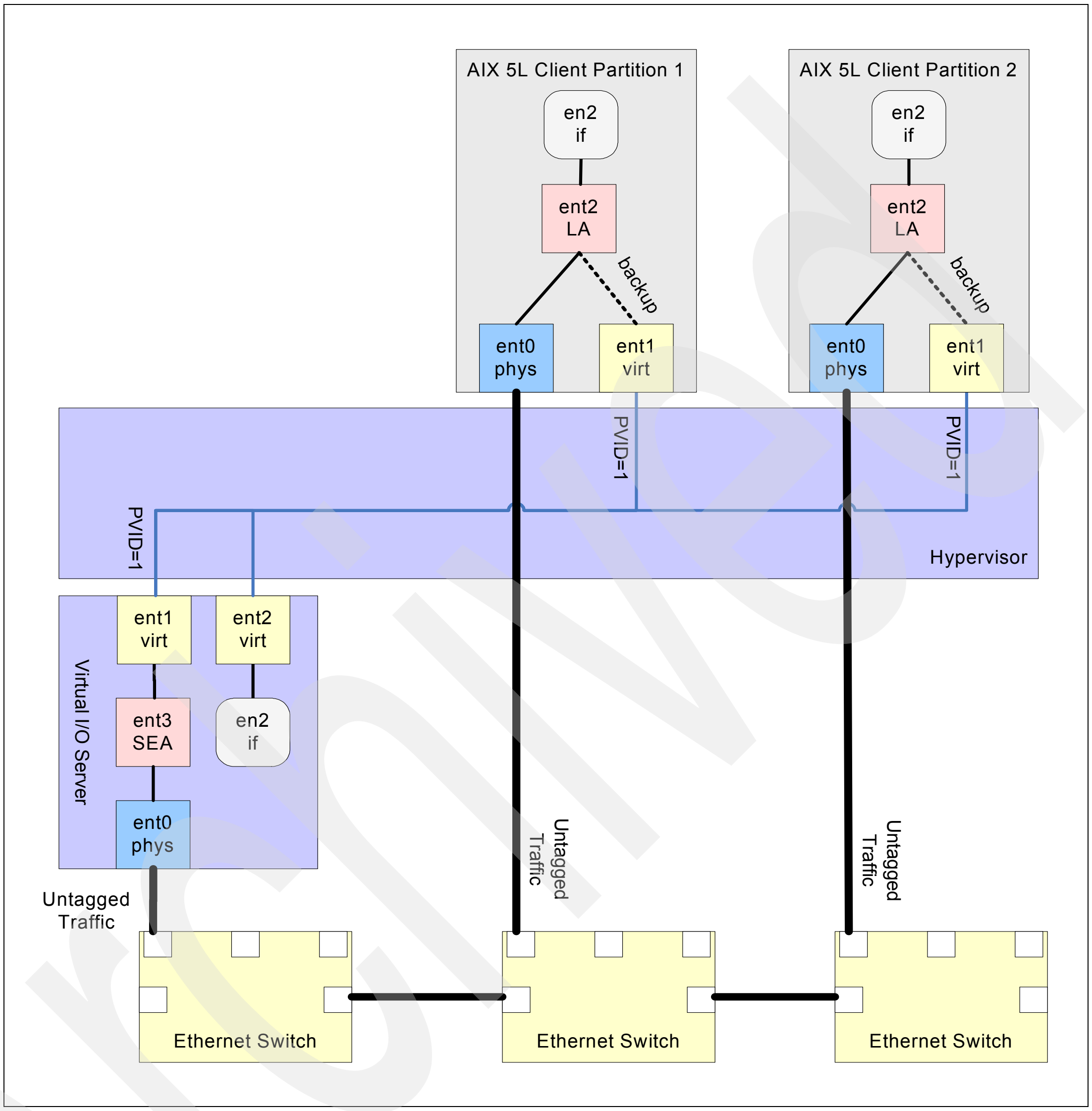


Figure 2-4 One Virtual I/O Server with client-partition backup link feature

Description of the scenario

In this scenario, client partitions use dedicated physical network adapters as their primary network connection. In order to provide some resilience if the physical connection suffers a network outage, the client is configured to use the backup channel of a link aggregation device. If a link-down event on the primary channel is detected, the link aggregation device then directs traffic to the backup channel, which in this case is a virtual path through a single Virtual I/O Server.

Description of the diagram

Figure 2-4 on page 26 shows that the Virtual I/O Server is configured in a standard manner with a single virtual network adapter, a single Shared Ethernet Adapter, and a single physical Ethernet adapter (the standard SEA configuration) for connection to the Ethernet switch. The physical Ethernet adapter of the Virtual I/O Server can be used as a backup by one or many client partitions where no access to the external network is readily available (for example, an Ethernet switch failure hosting several physical client connections).

This scenario provides resilience for client network connections and can protect them from network failures if the network connectivity has been designed with this in mind.

> **Note:** Because the Virtual I/O Server implements a Shared Ethernet Adapter, it is important to consider the total network load under certain network path failures. For example, if many client partitions suffer a link-down event due to an Ethernet switch failure, the traffic from those clients will share the single physical adapter in the Virtual I/O Server.
>
> Some activities, such as the network load generated with a network-based backup, also need to be taken into consideration when considering the number of network adapters in a configuration.

A number of variants to this basic scenario are possible, each offering different advantages. To provide more network throughput capability by the Virtual I/O Server, two physical adapters can be used with link aggregation. This enables more network traffic to be processed in the event of multiple clients requiring the use of the connection simultaneously. Alternatively, an additional SEA can be configured within the Virtual I/O Server with some clients targeting one SEA, and the rest targeting the other.

Architectural attributes

Table 2-3 provides the architectural attributes.

Table 2-3 One Virtual I/O Server with client-partition backup feature

	Network access
Resilience	Medium
Serviceability	High
Scalability/throughput	Standard
Flexibility	Medium

The following points explain the attributes and their rankings:

- Resilience

  Medium
  The configuration offers resilience above standard and protects client partitions from various simple network failures. This particular example design does not offer full redundancy because the throughput on the SEA is limited to that of the single physical Ethernet adapter. Therefore, consideration must be given to the circumstances where several client network connections sustain a failure, and the total aggregated traffic directed through the Virtual I/O Server might exceed that of a single adapter.

- Serviceability

  High
  This configuration has one Virtual I/O Server. It might be necessary to perform maintenance while the Virtual I/O Server is offline. During these short periods, the client partitions will be without the Virtual I/O Server capabilities. In this scenario, the Virtual I/O Server will only be carrying traffic if a network failure occurs that affects the physical adapter in a client partition. Therefore, under normal conditions where no traffic is carried through the Virtual I/O Server, the Virtual I/O Server can be maintained without any impact.

- Scalability/throughput

  Standard
  Many additional client partitions can be created. However, consideration must be given to the aggregated traffic throughput if a network failure occurs because the backup channel uses a single physical network adapter hosted by the Virtual I/O Server.

  All partitions share throughput of one physical Ethernet adapter in the Virtual I/O Server when a network failure occurs. The backup connection sustains network access for the aggregated client-partition traffic load, up to the limit of the one physical adapter.

- Flexibility

  Medium
  Client partitions can be added, but each will require its own physical Ethernet adapter for its primary network connection.

Summary and applicable workloads

This configuration provides network resilience for the client partitions. It is suited for non-complex Ethernet backup situations where there is value in protection from low-impact network failures such as an individual port or cable failure.

This scenario is suitable for solutions with the following characteristics:

- There is requirement for network resilience.
- There is no restriction of the resources such as physical Ethernet adapters and network switches.
- Test, development, and production environments.
- Web, application, database, and file servers.

Solutions with the following characteristic are not recommended:

- Network throughput of all the client partitions must be guaranteed over the backup connection on the Virtual I/O Server.

Further reference

For background and explanatory reading on technologies mentioned in this scenario, see the following references:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768
- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *Partitioning Implementations for IBM eServer p5 Servers*, SG24-7039

2.1.5 One Virtual I/O Server with link aggregation and backup link: Scenario 4

Figure 2-5 shows the Advanced POWER Virtualization architecture using a single Virtual I/O Server with a link aggregation (LA) device in backup mode for physical network access.

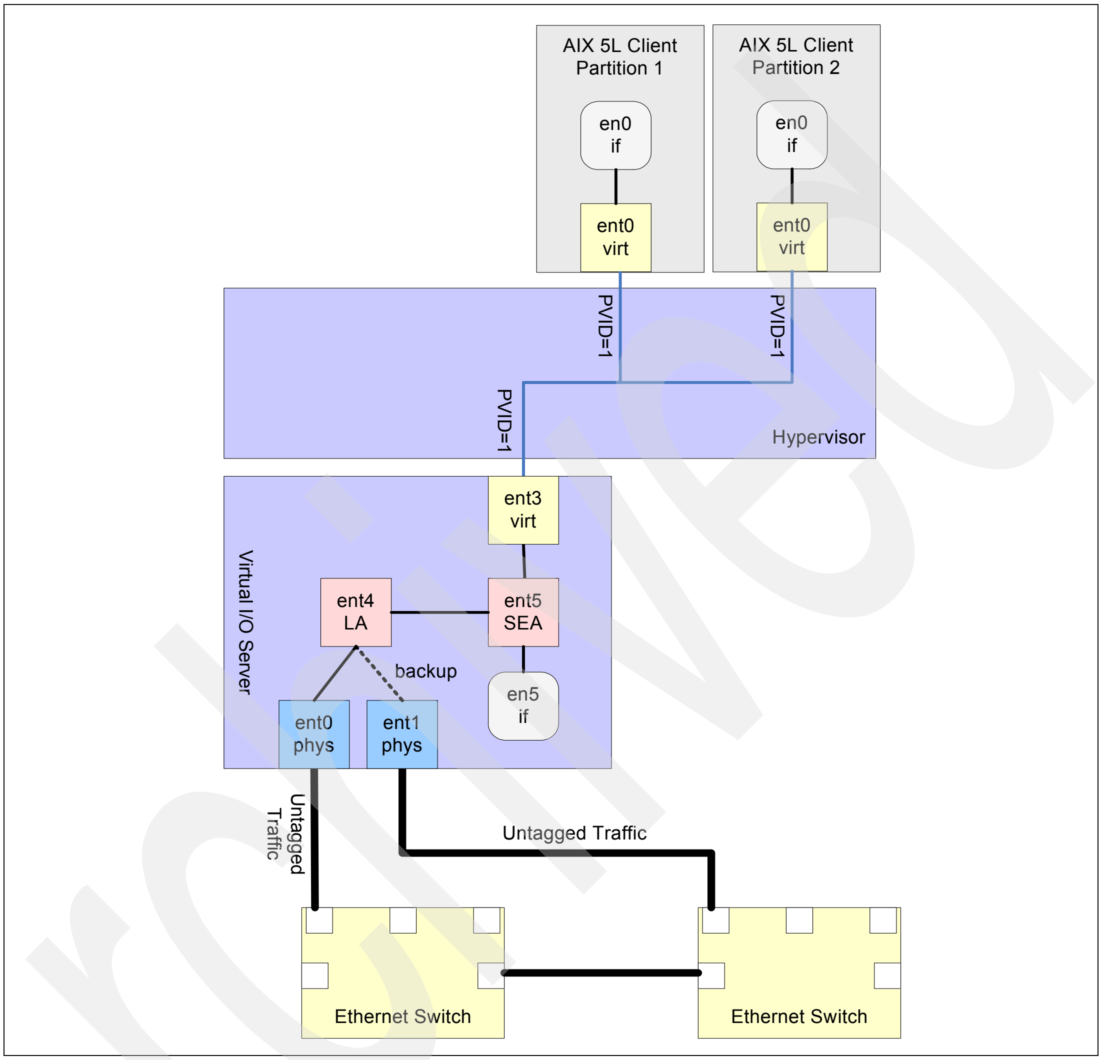


Figure 2-5 One Virtual I/O Server with link aggregation (LA) device in backup mode

The following figure shows an extension to simple backup mode, where the Advanced POWER Virtualization architecture uses single Virtual I/O Server with a link aggregation (LA) device in load balancing mode with a backup link (Figure 2-6).

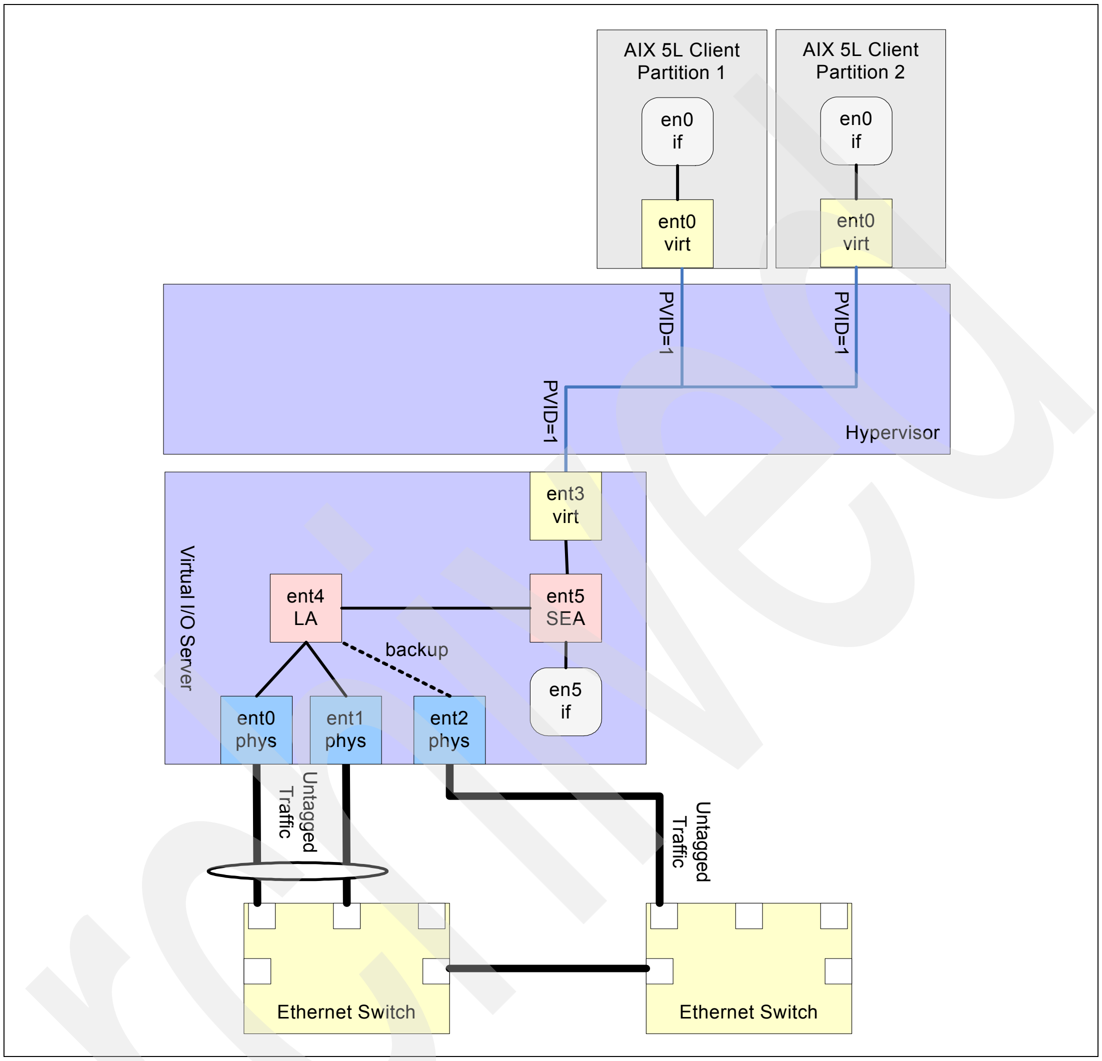


Figure 2-6 One Virtual I/O Server with LA device in load balancing mode with backup link

Description of scenario

These two variant scenarios provide increased resiliency of the connection to the external network. They use a single Virtual I/O Server with multiple links to the external network. They protect the client partitions if either a physical Ethernet adapter or a network switch becomes unavailable. They also have the potential to provide high throughput.

The Shared Ethernet Adapter in the Virtual I/O Sever is configured on the link aggregation device (LA) instead of individual physical adapters.

The link aggregation device with backup link is a software feature of AIX 5L. It has two links: a primary and backup. The backup link is optional. The primary link can operate either in EtherChannel mode or IEEE 802.3ad mode and consists of one (Figure 2-5 on page 29) or multiple (Figure 2-6) adapters. Network traffic is rerouted to the backup link only if all the network connections in the primary link fail. The solution is scalable, because additional network adapters can be added using dynamic LPAR operations and hot-plug tasks and incorporated into the LA device. For more information about LA technology, refer to “Overview of the link aggregation technology and features” on page 16.

With a backup link, no modifications are needed to the network settings (such as routing) and it can be connected through redundant network hardware.

If there is one physical adapter in the primary link, this mode of operation is referred to as *backup mode* (Figure 2-5 on page 29). The solution provides network redundancy but does not provide network load sharing. With this technology, the client partitions are protected from a single physical Ethernet adapter, cable, and switch failure. However, the network bandwidth to the external network for all client partitions is limited to one physical Ethernet adapter. To improve network throughput, consider the following recommendations:

- Allocate a dedicated physical network adapter to the client partitions that have to communicate very frequently and in high volumes to the external network (see 2.1.4 “One Virtual I/O Server with client-partition backup link: Scenario 3” on page 26).
- Use multiple active physical links under the link aggregation device in the Virtual I/O Server. There are two ways using AIX 5L to accomplish this: EtherChannel (Cisco proprietary protocol) or IEEE 802.3ad standard protocol. EtherChannel mode requires a Cisco switch and appropriate port switch configuration (not shown in our scenarios). IEEE 802.3ad mode (as presented in Figure 2-6 on page 30) is able to negotiate the protocol automatically with compatible switches. We use IEEE 802.3ad in this scenario to show the load balancing functionality.
- Use network separation with separate VLAN IDs and multiple physical and shared adapters in the Virtual I/O Server. In 2.1.4 “One Virtual I/O Server with client-partition backup link: Scenario 3” on page 26, we show this example scenario.

If high throughput and the use of multiple adapters is required, the basic scenario (Figure 2-5 on page 29) can be extended with *load balancing* in the LA device (Figure 2-6 on page 30). The link aggregation device can spread the network traffic between two or more physical adapters that make up the primary link (this mode of operation is defined by the IEEE 802.3ad standard). It provides scalable network bandwidth by using aggregated bandwidth of the adapters.

The scenario requires an appropriate network topology. Although it provides inherent protection against adapter and cable failures, there is still a possibility that the whole network switch might became unavailable. To protect from this situation, the backup link must be connected to a different network switch from the adapters in the primary link. This results in the client partitions still being able to communicate with the external network regardless of a network switch failure.

Description of one VIOS with LA device in backup mode (Figure 2-5 on page 29)

In Figure 2-5 on page 29, the SEA (ent5) is configured on ent3 and ent4. Ent3 is the Virtual I/O Server’s virtual Ethernet adapter. Ent4 is the pseudo-adapter that is configured as the LA with a backup link device on the physical Ethernet adapters ent0 and ent1. Because the virtual Ethernet adapters in the Virtual I/O Server and the client partitions must belong to the same VLAN in order to communicate, the VLAN ID is the same (default VLAN ID=1) for all the client partitions and the Virtual I/O Server’s virtual network adapters.

Under normal conditions, when ent0 on the Virtual I/O Server is the active physical Ethernet adapter, the client partitions will communicate to the external network using ent0. But in the case of ent0 failure, the network packets from the client partitions are routed to ent1 on the Virtual I/O Server. Therefore, the client partitions are still able to communicate. If the physical adapter providing the backup link is connected to a separate switch (distinct from the primary link), this scenario provides protection for a network switch outage.

Description of one VIOS with LA in load balancing mode (Figure 2-6 on page 30)

In Figure 2-6 on page 30, the Shared Ethernet Adapter (ent5) is configured on ent3 and ent4. Ent3 is the Virtual I/O Server's virtual Ethernet adapter. Ent4 is the pseudo-adapter that is configured as the link aggregation device with multiple physical Ethernet adapters making up the primary link (ent0 and ent1) in the load balancing mode and backup link (ent2). In normal operation, the ent0 and ent1 adapters provide the active link, and ent2 is the backup (standby) adapter. The Ethernet adapters (ent0, ent1) in the active mode must be plugged into the same network switch, and special configuration might be required in the network switch. For the backup Ethernet adapter (ent2), there is no special configuration in the network switch required.

> **Tip:** For the link aggregation configuration to spread network traffic among multiple physical Ethernet adapters, the network switch must be configured appropriately.

If both active Ethernet adapters (ent0 and ent1) are not able to communicate with the switch, traffic is automatically rerouted to the backup Ethernet adapter (ent2) and the alternate network switch. This link aggregation configuration protects the client partitions from a failure in a network switch, cabling, and physical Ethernet adapter. However, during the time the backup link is active, the network bandwidth is limited to a single physical Ethernet adapter.

Architectural attributes

In this scenario, two diagrams are shown. They present basically the same functionality but they have different ratings in the architectural attributes tables (the link balancing solution is more scalable and provides higher throughput). The tables and their text descriptions for the individual diagrams follow.

One Virtual I/O Server with link aggregation device in backup mode

Table 2-4 provides the architectural attributes.

Table 2-4 One VIOS with LA device in backup mode (Figure 2-5 on page 29)

	Network access	
Resilience	Medium	🟩🟩
Serviceability	Standard	🟩
Scalability/throughput	Standard	🟩
Flexibility	High	🟩🟩🟩

The following points explain the attributes and their rankings:

- Resilience

  Medium
  Protects against a single switch and a single Ethernet adapter failure, but maintenance or unplanned unavailability of the whole Virtual I/O Server will block the communication to external network.

- Serviceability

  Standard
  Maintenance on network infrastructure (adapters, cabling, and switch, for example) can be done without impacting the client partitions, but the client partitions will not be able to communicate to external network during maintenance of the Virtual I/O Server itself.

- Scalability/throughput

  Standard
  The client partitions share one physical adapter for external communication. The total aggregated throughput is limited to the performance of one adapter.

- Flexibility

  High
  VLANs can be added easily to both Virtual I/O Server and client partitions by using dynamic LPAR operations.

  Client partitions are easily added on demand without additional work to install hardware if there are available resources such as CPU, memory, and physical Ethernet adapter throughput.

One Virtual I/O Server with LA in load balancing mode with backup link

Table 2-5 provides the architectural attributes of this configuration.

Table 2-5 One VIOS with LA in load balancing mode with backup link (Figure 2-6 on page 30)

	Network access	
Resilience	Medium	🟩🟩
Serviceability	Standard	🟩
Scalability/throughput	Medium	🟩🟩
Flexibility	High	🟩🟩🟩

The following points explain the attributes and their rankings:

- Resilience

  Medium
  Protects against a switch and a Ethernet adapter failure, but maintenance or unplanned unavailability of the whole Virtual I/O Server will block communication to external networks.

- Serviceability

  Standard
  Maintenance on network infrastructure (adapters, cabling, and switch, for example) can be done without impacting the client partitions, but the client partitions will not be able to communicate to the external network during maintenance of the Virtual I/O Server itself.

- Scalability/throughput

  Medium
  Additional clients can be easily added to scale up the configuration without additional hardware installation as long as the additional client partitions do not exceed the network bandwidth of the adapters in the primary link of the LA device.

  Load balancing in the LA device allows for network load sharing between multiple physical Ethernet adapters, giving the clients higher network bandwidth. However, if all the primary adapters fail, the network bandwidth decreases to the throughput of one adapter.

- Flexibility

  High
  VLANs can be added easily to both the Virtual I/O Server and client partitions by using dynamic LPAR operations.

Client partitions can be easily added on demand without additional work to install hardware if there are available resources such as CPU, memory, and physical Ethernet adapter throughput.

Adding physical network adapters to the Virtual I/O Server requires a hardware change, but the associated additional link aggregation configuration is easy and will not interrupt production.

Summary and applicable workloads

In these two variations of the scenario, higher network resiliency to the external networks is implemented with LA technology and two switches (Figure 2-5 on page 29). In addition, higher network bandwidth can be provided by using multiple physical Ethernet adapters as the active link with the LA device (Figure 2-6 on page 30).

This scenario can be applicable to many workloads where access to the external network and resilience to various network conditions is required. It protects from all types of network failures, except the case of unavailability of the Virtual I/O Server itself. The scenario can be used for a range of applications, from low network throughput to very high network throughput, depending on workload characteristics.

This configuration is suitable for solutions with the following characteristics:

- High volume of network traffic between the client partitions and moderate collective network traffic to the external networks (Figure 2-5 on page 29).
- High volume of collective network traffic to the external network (Figure 2-6 on page 30).
- External network communication is critical but planned downtime is allowable.
- Dual Virtual I/O Servers are not desired.
- Test, develop, and production environments that allow service windows.
- For web, application, and database servers.

Solutions with the following characteristics are not recommended:

- Network downtime is not acceptable.
- If most of network traffic to external network comes from only one client partition and the network traffic severely affects the total network performance, we recommend a dedicated physical Ethernet adapter.
- A production environment that requires high network serviceability.

Further reference

For more detailed information for this scenario, see the following references:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768

2.1.6 Dual Virtual I/O Servers with client LA backup channel: Scenario 5

Figure 2-7 outlines the basic network configuration for this scenario.

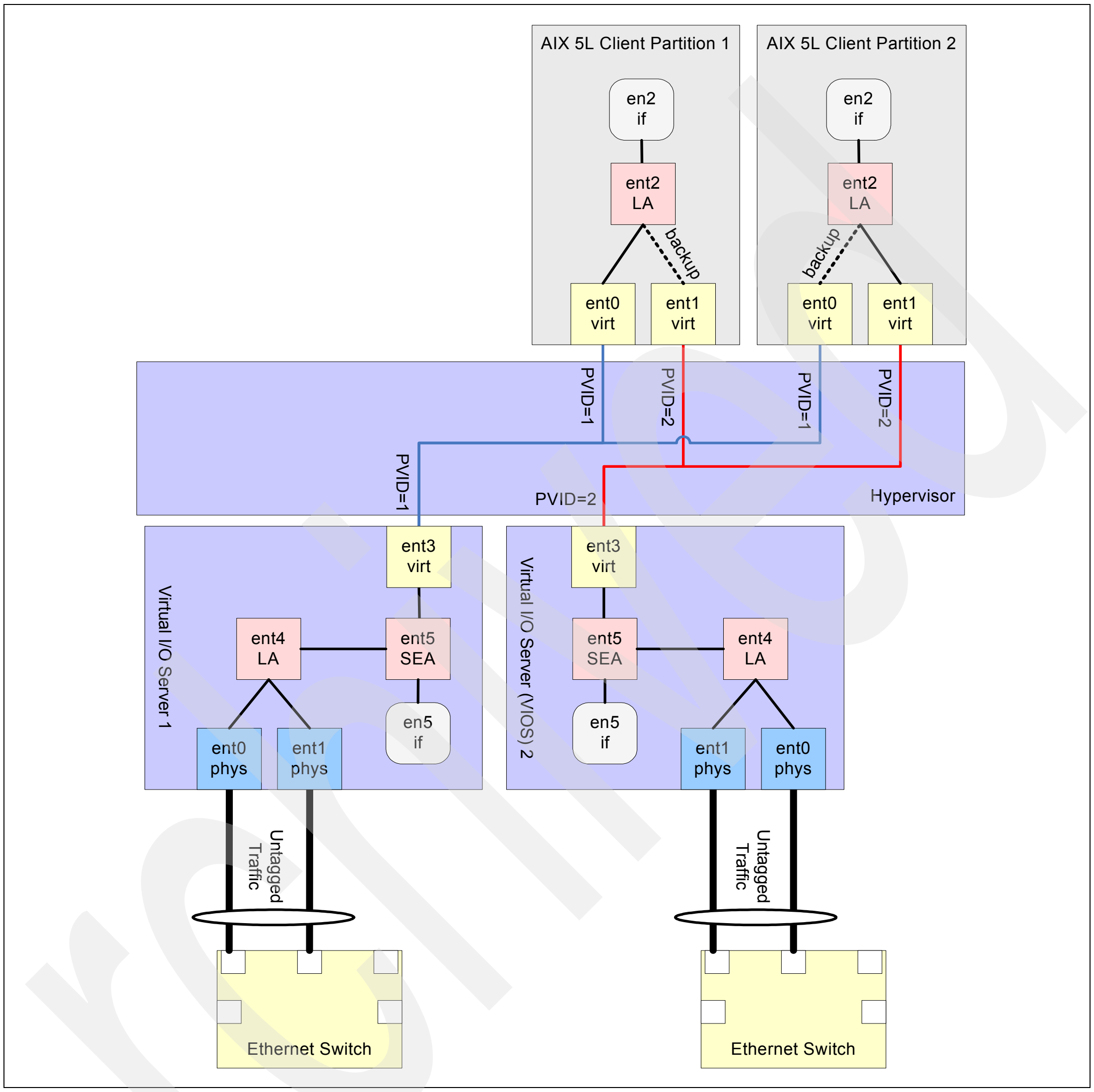


Figure 2-7 Dual Virtual I/O Servers with client LA backup channel

Description of scenario

With a single Virtual I/O Server, client partitions have the opportunity to improve redundancy (and thus become more resilient) by implementing a second Virtual I/O Server. The client partitions can exploit the two Virtual I/O Servers in a variety of ways.

In this scenario, the client partitions use the Virtual I/O Servers to implement a backup connection (link aggregation with backup channel). Each client partition has a primary network connection with a path to one of the Virtual I/O Servers and out to the physical network, as well as a backup connection with a path to the alternate Virtual I/O Server. The backup connection is activated only if the primary connection fails. In this way, a highly available, resilient network topology can be deployed.

This scenario provides the concepts for increasing the resilience and serviceability of the Virtual I/O Server and for providing the shared Ethernet service to client partitions. However, it is important to realize that deploying two Virtual I/O Servers will require additional hardware resources.

> **Note:** Virtual Ethernet adapters do not support link aggregation technology such as EtherChannel and IEEE 802.3ad in load balancing mode. But link aggregation with the backup feature can be configured on virtual Ethernet adapters.

Advantages

Network traffic load can be shared between Virtual I/O Servers.

By controlling which virtual Ethernet adapter becomes the active adapter, network traffic can be routed to a specific Virtual I/O Server. Therefore, the network load can be shared between dual Virtual I/O Servers.

SEA failover operates in primary-backup mode. This means that client partitions can use the backup SEA only when primary SEA becomes unavailable. Therefore, all the network load is concentrated on the primary SEA all the time, which might lead to network bottlenecks.

Description of the diagram

As shown in Figure 2-7 on page 35, the dual Virtual I/O Servers have a Shared Ethernet Adapter to bridge to the external network. Each Shared Ethernet Adapter on both Virtual I/O Servers connects to a different VLAN. Virtual I/O Server 1 connects to VLAN 1 (default VLAN ID=1) and Virtual I/O Server 2 connects to VLAN 2 (default VLAN ID=2). Each client partition has two virtual adapters, each adapter connecting to a different VLAN (with a different default VLAN ID).

For network resilience, the link aggregation with backup feature is used on the client partitions in this scenario. The link aggregation with backup channel feature described in “Overview of the link aggregation technology and features” on page 16 is a technology that provides network redundancy through the use of an active channel and backup channel. Each client partition has its IP address defined on the pseudo-network adapter (ent4) that is the link aggregation device. The link aggregation device is configured on the two virtual Ethernet adapters (ent0 and ent1) that are connected to different VLANs.

The backup channel configuration in the client partitions provides high network resilience for the client partitions exploiting the dual Virtual I/O Servers. The backup channel is inactive (standby) during normal operation, meaning that only one virtual Ethernet adapter on a client partition can be active and communicate through one Virtual I/O Server at any time.

As shown in Figure 2-7 on page 35, client partition 1 can communicate using ent0 to VLAN1, which is connected to Virtual I/O Server 1 for its primary (active) connection to the network. However, if Virtual I/O Server 1 requires intervention, the backup channel is activated (ent1 in VLAN2 becomes active). Therefore, client partition 1 can remain connected to the external network.

Client partition 2 uses ent1 in VLAN2 to connect to Virtual I/O Server 2 for its primary (active) connection to the network. Similarly, if Virtual I/O Server 2 requires intervention, client partition 2 activates the backup channel and uses ent0 with Virtual I/O Server 1 to connect to external network.

For a scenario such as this, the primary (active) adapter in the client partition must poll (ping) or test for network connectivity using an external IP address on a reachable network to validate the traffic path. If the traffic path is not valid for some reason (the ping fails), the backup channel in the client partition is activated and traffic will flow through the backup adapter. This test IP address is only valid when using a backup adapter, and configuring a test IP address is mandatory for a backup channel configuration. This is because no link-down event can be detected in a virtual environment.

Each of the client network adapters (primary and backup) must be connected to different Virtual I/O Servers with different VLANs. Additional VLAN definitions through a single SEA (implementing multiple VLANs) are not permitted in this scenario.

> **Note:** When using a link aggregation with a backup channel configuration for a virtual Ethernet adapter, the external IP address configuration to verify network connectivity is mandatory. Two virtual Ethernet adapters must use different VLAN IDs that are reflected in the default VLAN IDs. Additional VLAN definitions (implementing multiple VLANs) are not permitted.

Network load distribution between the Virtual I/O Servers

Network traffic load can be distributed between the two Virtual I/O Servers by controlling which client adapter (active and backup adapters) in each client partition connects to each of the Virtual I/O Servers. This approach allows network traffic to be shared between both Virtual I/O Servers.

For example, in Figure 2-7 on page 35, the active adapter for client partition 1 is ent0 in VLAN1, and the active adapter for client partition 2 is ent1 in VLAN2. In this case, the network traffic of client partition 1 will be processed by Virtual I/O Server 1, and the network traffic of client partition 2 will be processed by Virtual I/O Server 2. Using this technique, network bottlenecks to the external network can be avoided and this also exploits the capabilities of a dual Virtual I/O Server configuration (the network traffic must be separated by a VLAN).

Redundancy of physical network attachment in a Virtual I/O Server

With dual Virtual I/O Servers and the link aggregation with backup adapter feature used in both client partitions, the network connectivity to the external network has increased resilience. However, even greater resilience can added by using link aggregation within the Virtual I/O Servers connecting to the Shared Ethernet Adapter.

By configuring link aggregation on the Shared Ethernet Adapters of both Virtual I/O Servers, the bandwidth limitations of a single physical Ethernet adapter can be overcome, avoiding any traffic bottleneck when sharing one physical Ethernet adapter among many client partitions.

For example in Figure 2-7 on page 35, during normal running conditions, ent0 and ent1 on Virtual I/O Server 1 are available and can fully use the capacity of both physical Ethernet adapters. If ent0 on Virtual I/O Server 1 becomes unavailable, the network packets are automatically sent through ent1 without disruption of existing client partition connections. It is important to remember that a link or adapter failure can cause a capacity degradation that might correspondingly affect network performance.

To achieve greater network bandwidth, additional physical Ethernet adapters can be added to an existing link aggregation dynamically using the Dynamic Adapter Membership (DAM) feature (available in AIX 5L). The network switch must be also configured properly to support the new adapters.

Architectural attributes

Table 2-6 provides the architectural attributes.

Table 2-6 Dual Virtual I/O Servers with backup

	Network access	
Resilience	High	🟩
Serviceability	High	🟩
Scalability/throughput	High	🟩
Flexibility	Medium	🟩

The following points explain the attributes and their rankings:

- Resilience

  High
  Protects against a single Virtual I/O Server, switch, and a single Ethernet adapter intervention.

- Serviceability

  High
  When one Virtual I/O Server is not available due to maintenance, the alternative Virtual I/O Server can provide all network connections.

- Scalability/throughput

  High
  With link aggregation on the Shared Ethernet Adapter, the network bandwidth to the external network is extended above that of one physical Ethernet adapter. Additional physical Ethernet adapters can be added to an existing link aggregation device to support more network traffic.

  Network traffic can be distributed between both Virtual I/O Servers by controlling which virtual Ethernet adapter will be the primary (active) on each client partition.

- Flexibility

  Medium
  Client partitions can be added easily without additional work to install hardware. VLANs can be added to both Virtual I/O Server and client partitions by using dynamic LPAR.

Summary and applicable workloads

This scenario provides the network redundancy by implementing link aggregation with the backup adapter feature on the client partitions with dual Virtual I/O Servers. This configuration also eliminates a client's communication disruption due to network switch or Virtual I/O Server maintenance. You can do planned maintenance while the alternate Virtual I/O Server and switch are still online.

Pertaining to throughput, this scenario provides these features:

- Controlling which virtual Ethernet adapter is active on client partitions means that client partitions can decide which Virtual I/O Server is used primarily. The network traffic from client partitions can be distributed between both Virtual I/O Servers. Therefore, it is possible to prevent SEA network bottlenecks.
- By combining the Shared Ethernet Adapter with the link aggregation feature, you can overcome the bandwidth limitation of a single network adapter and avoid bottlenecks.

- For additional network bandwidth, additional physical Ethernet adapters can be dynamically added to an existing link aggregation adapter.

However, the overall complexity is greater than with a single Virtual I/O Server:

- All the client partitions have to be configured with link aggregation with the backup adapter to increase network resilience. Consider the distribution of network traffic between both Virtual I/O Servers.
- The link aggregation configuration for Shared Ethernet Adapters and network switches is required on both Virtual I/O Servers.

This scenario can match many workloads where highly available access to the external network and concurrent maintenance of Virtual I/O Server is required. This scenario can be used for range from low network throughput to very high network throughput depending on workload characteristics:

This configuration is suitable for solutions with the following characteristics:

- External network communication is critical.
- Network serviceability and resilience are very important.
- A lot of network traffic to the external network.
- The network traffic from client partitions can be predicted, and it is possible to plan distributing network traffic between both Virtual I/O Servers.
- Dual Virtual I/O Servers can be managed.
- There is the need to update Virtual I/O Server software.
- Many environments such as test, development, and production.
- Application server and database server, and especially Web server traffic can be distributed between both Virtual I/O Servers.

Solutions with the following characteristics are not recommended:

- Most of network traffic to external network comes from only one client partition, and the network traffic from a client partition severely affects the total network performance.
- The file server generates a large amount of network traffic, which affects the overall network throughput.

Further reference

For more detailed information for this scenario, see the following references:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768

2.1.7 Dual Virtual I/O Servers with SEA failover: Scenario 6

Figure 2-8 outlines the basic architecture for this scenario.

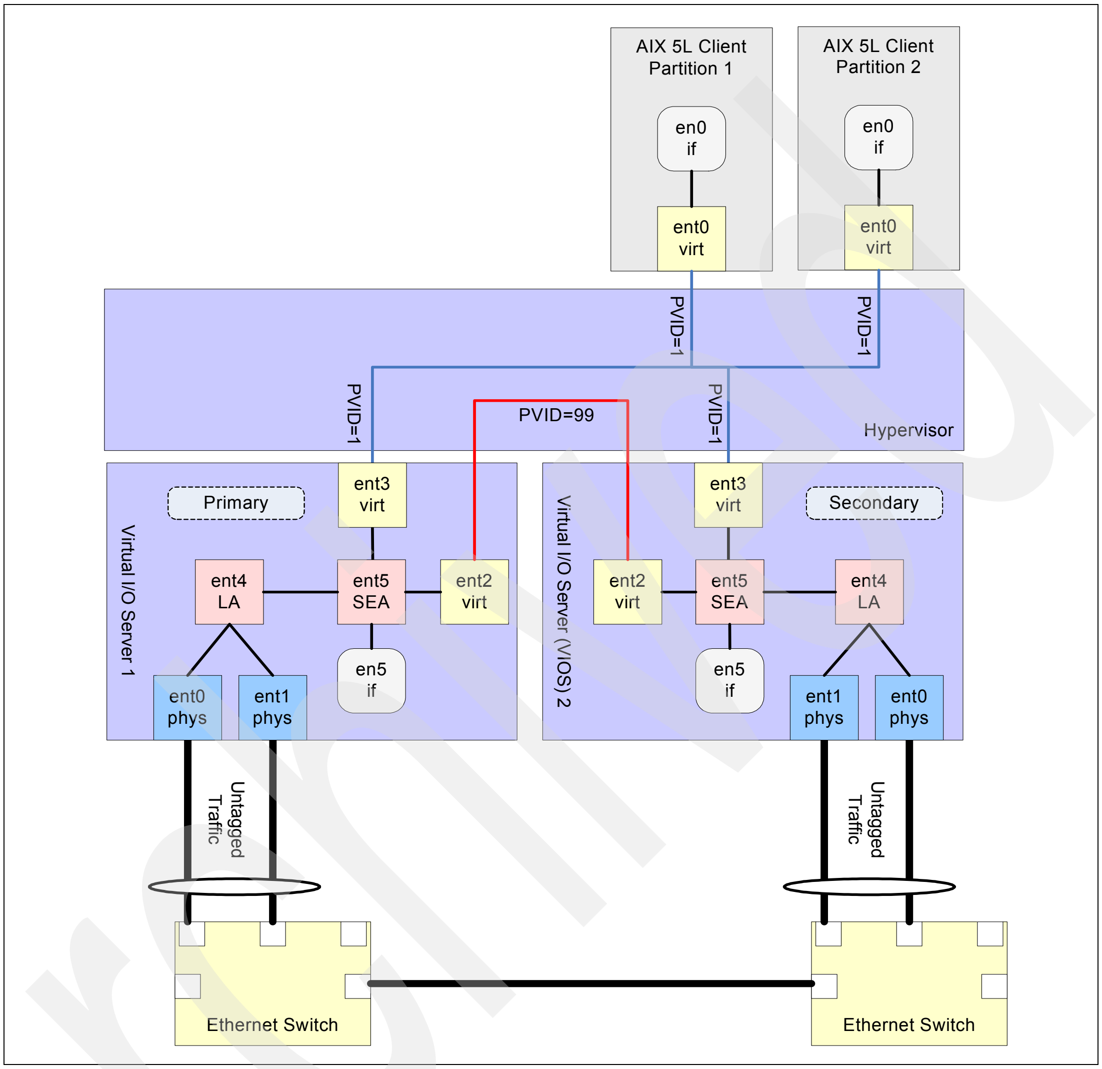


Figure 2-8 Dual Virtual I/O Servers with SEA failover

Description of scenario

With the previous scenario, 2.1.6 “Dual Virtual I/O Servers with client LA backup channel: Scenario 5” on page 35, client partitions are protected from intervention required on a Virtual I/O Server, a physical Ethernet adapter, and a network switch.

In this scenario, it is possible to implement the same network redundancy by using the Shared Ethernet Adapter failover feature on the Virtual I/O Servers (new in Virtual I/O Server 1.2). All the client partitions within the system are provided with high resilience without any special configuration.

Because there are similarities in functionality between this scenario and that of scenario 5, we outline the strengths of this scenario.

The main strengths of this scenario are:

- Additional VLAN IDs available with a VLAN-aware network switch.

  If there is a VLAN-aware network switch, there is no restriction about adding a new VLAN ID to an existing virtual Ethernet adapter. Similarly, there is no restriction about adding a new virtual Ethernet adapter with a newly configured VLAN ID to the SEA in this scenario. Neither of these is possible in scenario 5.

- Less complex client configurations.

  Only one virtual Ethernet adapter is required in the client partitions and there is no special configuration. In contrast to this, for scenario 5, the LA device must be configured in the client partitions on a pair of virtual Ethernet adapters.

- Network link detection happens in Virtual I/O Server.

  The clients do not have to ping a network address to verify the network status.

Description of the diagram

In Figure 2-8 on page 40, the dual Virtual I/O Servers use the Shared Ethernet Adapter failover feature that offers network redundancy to the client partitions at the virtual level. The client partitions have one standard virtual Ethernet adapter. The external traffic is handled by two Virtual I/O Servers and they use a control channel to control the failover, and thus, provide uninterrupted service to the client partitions. The client partitions do not need to have special software or special configurations for their virtual Ethernet adapters.

As shown in Figure 2-8 on page 40, both Virtual I/O Servers attach to the same VLAN (default VLAN ID 1). Both virtual Ethernet adapters on both Shared Ethernet Adapters have the access the external network option in the Virtual I/O Server's profile enabled with trunk priority option that decides which Shared Ethernet Adapter will be the primary.

An additional virtual Ethernet connection has to be set up as a separate dedicated VLAN (default VLAN ID 99) between the two Virtual I/O Servers and must be attached to the Shared Ethernet Adapter. This connection is a control channel to exchange keep-alive or heartbeat messages between the two Virtual I/O Servers and thus controls the failover of the bridging functionality. Apart from the control channel adapters, no other device is attached to the control channel.

In addition, the Shared Ethernet Adapters failover feature can detect certain network failures by periodically pinging the IP address that is configured on the Shared Ethernet Adapter to confirm the network connectivity.

> **Note:** The following points are of special consideration:
>
> - You must configure a control channel for the SEA failover feature.
> - A lower number trunk priority has a higher scheduling priority.
> - Do not confuse the IP address used in the SEA for the external network ping availability test with the IP address used to access the Virtual I/O Server for administration purposes.

The Shared Ethernet Adapter failover is initiated by:

- The standby Shared Ethernet Adapter detects that keep-alive messages from the active Shared Ethernet Adapter are no longer received over the control channel.
- The active Shared Ethernet Adapter detects that a loss of the physical link is reported by the physical Ethernet adapter's device driver.

- On the Virtual I/O Server with the active Shared Ethernet Adapter, a manual failover can be initiated by setting the active Shared Ethernet Adapter to standby mode.
- The active Shared Ethernet Adapter detects that it cannot ping a given IP address.

> **Note:** Shared Ethernet failover is only supported on the Advanced POWER Virtualization Virtual I/O Server V1.2 and later.

Redundancy of physical network attachment in Virtual I/O Server

For the redundancy of the Shared Ethernet Adapter itself, consider link aggregation, such as EtherChannel or IEEE 802.3ad. Link aggregation is a network port aggregation technology that allows several network adapters to be aggregated together into a single pseudo-network adapter. All physical Ethernet adapters that consist of the same link aggregation must be connected to the same network switch, and additional configuration work might be required in the network switch.

This technology is often used to overcome the bandwidth limitation of a single physical Ethernet adapter and avoid bottlenecks when sharing one physical Ethernet adapter among many client partitions. If one network adapter fails, the network packets are automatically sent through another available network adapter without disruption of existing user connections. However, a link or adapter failure might lead to performance degradation.

In this scenario, each SEA has two physical Ethernet adapters and two virtual Ethernet adapters. As mentioned previously, one of the virtual Ethernet adapters is for the control channel to monitor and the other is for communication between the Virtual I/O Server and client partitions. The two physical Ethernet adapters are connected to the same external network switch with the link aggregation configuration. All these Ethernet adapters are configured to a Shared Ethernet Adapter that bridges network packets from client partitions to the external network and vice versa.

With link aggregation, you can overcome the bandwidth limitation of a single physical Ethernet adapter using additional physical Ethernet adapters. Additional adapters can be added to an existing link aggregation using the Dynamic Adapter Membership (DAM) feature supported by AIX 5L. At this time, the physical network switch must also be configured to support the newly connected physical adapter.

> **Note:** A virtual Ethernet adapter does not support link aggregation such as EtherChannel or IEEE 802.3ad.

Architectural attributes

Table 2-7 provides the architectural attributes.

Table 2-7 Dual Virtual I/O Servers with SEA failover

	Network access	
Resilience	High	🟩
Serviceability	High	🟩
Scalability/throughput	High	🟩
Flexibility	High	🟩

The following points explain the attributes and their rankings:

- Resilience

  High
  Protects against a single Virtual I/O Server, switch, a single Ethernet adapter, and cabling interventions.

- Serviceability

  High
  If one of the Virtual I/O Servers is taken offline for maintenance, an alternate network path to the external network is available through the remaining Virtual I/O Server. Similarly, this scenario can facilitate uninterrupted service during switch maintenance.

- Scalability/throughput

  High
  Client partitions can be added easily without additional work to install hardware if there are available resources such as CPU and memory.

  With the link aggregation configuration on the Shared Ethernet Adapter, the network bandwidth to the external network can be increased beyond one physical Ethernet adapter. Additional physical Ethernet adapters can also be added to an existing link aggregation to support more network traffic.

- Flexibility

  High
  VLANs with different VLAN IDs can be added to both Virtual I/O Servers and client partitions by using dynamic LPAR, and inter-partition communication can be enabled easily without the need for additional cabling work.

  This is more complex because additional link aggregation is required on both the Virtual I/O Servers and the network switches. Some specialized setup might be required to the Virtual I/O Server for failover Shared Ethernet Adapters.

Summary and applicable workloads

This scenario provides network resilience by implementing the Shared Ethernet Adapter failover feature on dual Virtual I/O Servers. Planned maintenance is possible while the alternate Virtual I/O Server and network switch are still online. In addition, by combining the Shared Ethernet Adapter with the link aggregation feature, it is possible to overcome the bandwidth limitation of a single network adapter.

It is easy to add another VLAN that communicates with the external network. The network switches must be VLAN-aware and configured appropriately.

However, the overall setup is slightly more complex than that of a single Virtual I/O Server. This is because of the special configurations required for both the SEA failover and the network switch (link aggregation feature).

This scenario can match many workloads where highly available access to the external network and concurrent maintenance of Virtual I/O Server are required. This scenario can be used where the requirements range from low network throughput to very high network throughput depending on workload characteristics.

The configuration is suitable for solutions with the following characteristics:

- External network communication is critical.
- Network serviceability and resilience are very important.
- A high volume of network traffic to the external network is expected.

- Each server requires network bandwidth at different times.
- Dual Virtual I/O Servers are manageable.
- Environments such as test, development, and production.
- Web, application, and database servers.

Solutions with the following characteristic are not recommended:

- Most of network traffic to external network comes from only one client partition, and the network traffic from a client partition severely affects the total network performance.

Further reference

For more detailed information for this scenario, see the following references:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768

2.2 Virtual I/O Server SCSI

The following section describes possible virtual SCSI scenarios. These scenarios are just a few of the many ways to design a Virtual I/O Server configuration to provide disk resources to the client partitions.

2.2.1 Introduction to virtual I/O SCSI terminology

Figure 2-9 outlines the basic virtual SCSI architecture and acts as the basis for the explanation of virtual SCSI terminology.

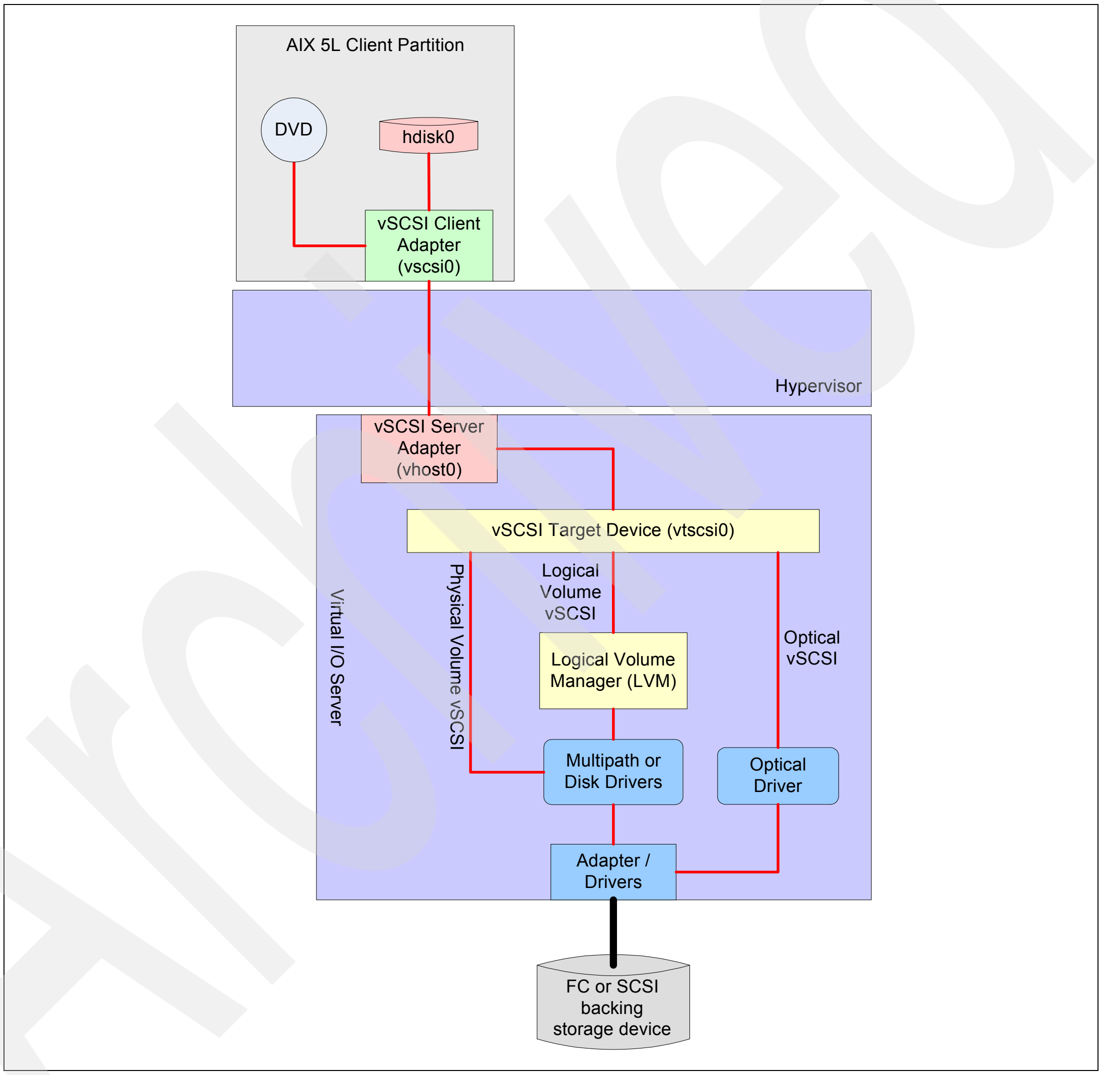

AIX 5L Client Partition
DVD
hdisk0
vSCSI Client Adapter (vscsi0)
Hypervisor
vSCSI Server Adapter (vhost0)
vSCSI Target Device (vtscsi0)
Virtual I/O Server
Physical Volume vSCSI
Logical Volume vSCSI
Logical Volume Manager (LVM)
Optical vSCSI
Multipath or Disk Drivers
Optical Driver
Adapter / Drivers
FC or SCSI backing storage device

Figure 2-9 Basic virtual SCSI client/server architecture

Virtual SCSI is the advanced technology that virtualizes the SCSI protocol. This is the technology that allows the sharing of physical adapters and disk devices among many client partitions.

Virtual SCSI is a client/server architecture where the Virtual I/O Sever services the storage needs of client partitions. The physical storage devices (which can also be referred to as backing storage devices) are attached to the Virtual I/O Server using physical storage adapters (for example, fibre channel, SCSI, or RAID adapters). The Virtual I/O Server defines logical or physical volumes that are exported to the client partitions.

On the virtual SCSI client partitions, these logical or physical volumes appear to be SCSI disks. To address its storage requirements, a client partition can use virtual SCSI storage, dedicated physical storage, or a combination of both.

> **Note:** The physical storage subsystem used as backing storage by virtual SCSI services needs to follow the normal performance and throughput design considerations, such as the capacity required and the I/O rate. An increase in the number of virtual SCSI clients will lead to a greater I/O load on the storage subsystem. It is important, therefore, to ensure that any virtual SCSI implementation is tested adequately so that it meets the required performance goals.
>
> When considering the physical storage subsystem used by the Virtual I/O Server as backing storage, take into account the following characteristics:
>
> - The number of disk spindles
> - The number of read/write arms on the disks
> - The rotational speed of the disks
> - The cache size on the storage device
> - The connection capabilities (for example, 2 Gbps FC)
> - Overall load on the storage subsystem from other workloads competing for service

Actual impact on overall system performance will vary by environment.

The Virtual I/O Server supports several types of storage backing devices to host the virtual SCSI disks, for example SAN, SCSI, and RAID devices. In addition, iSCSI is supported by the Virtual I/O Server in V1.3.

Virtual SCSI terminology

Note the following virtual SCSI terms:

Virtual I/O Server The server partition that provides virtual SCSI service to client partitions and hosts all server-side virtual SCSI adapters (one server-side virtual SCSI adapter is required for each connecting client partition path).

Virtual SCSI client The client partition that hosts a client-side virtual SCSI adapter and uses the physical or logical volumes provided by the Virtual I/O Server.

vSCSI server adapter The server-side virtual SCSI adapter hosted by the Virtual I/O Server provides virtual SCSI services.

vhost0 The device label used on the scenario diagrams for the server-side SCSI adapter hosted by the Virtual I/O Server.

vSCSI client adapter The client-side virtual SCSI adapter hosted by the client partition through which virtual SCSI services are accessed (such as physical or logical volumes).

vscsi0 The device label on the scenario diagrams for the client-side virtual SCSI adapter device.

vSCSI target device A virtual device instance that can be considered as a mapping device. It is not a physical device, but rather a mechanism for managing the mapping of the portion of physical storage (hosted by the Virtual I/O Server) that is presented to the client partition as a SCSI device.

vtscsi0 The device label on the scenario diagrams for the vSCSI target device.

Optical vSCSI The Virtual I/O Server supports the exporting of physical optical devices, such as CD-ROM, DVD-ROM, and DVD-RAM. These are called virtual SCSI optical devices and always appear as SCSI devices to client partitions regardless of the underlying infrastructure (be it SCSI, IDE, or USB).

Multipath Technology where one disk device or LUN can be accessed through multiple device adapters simultaneously. This is generally used for availability and performance reasons.

SAN environment LUN identification

In a flexible SAN environment, there is a need to unique identify LUNs based on some information other than their location codes. The location can change over time, and in addition to that, in multipath environments, each LUN is accessible through several adapters (thus having several location codes).

The old way of disk identification is based on the physical volume identifier. A disadvantage of this approach is that the physical volume identifier has to be written to the data space of the disk. Therefore, in the Virtual I/O Server environment, the beginning of the disk presented to the client partitions is shifted from the physical beginning to preserve space for the physical volume identifier written by the Virtual I/O Server. Data on such a disk cannot be accessed by AIX 5L if the LUN is moved to direct attachment. A change to direct-attach access requires backing up and restoring the data.

The preferred approach is to identify LUNs based on the unique disk identifier (UDID) that becomes standard in many SAN-attached devices. The UDID is generated by hardware and is not written in the data space. In order to use UDID identification in the Virtual I/O Server, appropriate multipath drivers compatible with UDID have to be used. When a Virtual I/O Server configures a LUN for the first time, it decides which approach for identification to use. It always prefers UDID if it is available for the given attached storage and supported by the installed driver. Therefore, it is preferred to use UDID compatible multipath software where available. For example, the Subsystem Device Driver Path Control Module (SDDPCM) (UDID compatible) is preferred over the Subsystem Device Driver (SDD).

Multipath device support

The following multipath devices are supported in Virtual I/O Server Version 1.3: AIX 5L multipath I/O (MPIO) (default PCM), SDDPCM, SDD, Redundant Disk Array Controller (RDAC), PowerPath (EMC), HDLM (Hitachi Data Systems), AutoPath (HP). The RDAC driver does not support load balancing. SDD currently does not support multipath access in client partitions. Third-party drivers might have additional configuration considerations. For detailed SAN-attached device support in the Virtual I/O Server, refer to:

http://techsupport.services.ibm.com/server/vios/documentation/datasheet.html

Further reference

For background and explanatory reading about these technologies, see the following references:

- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768
- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *Partitioning Implementations for IBM eServer p5 Servers*, SG24-7039
- *IBM System p5 Approaches to 24x7 Availability Including AIX 5L*, SG24-7196

2.2.2 One Virtual I/O Server with single path to storage: Scenario 7

Figure 2-10 depicts three variations of the basic disk configuration for this scenario.

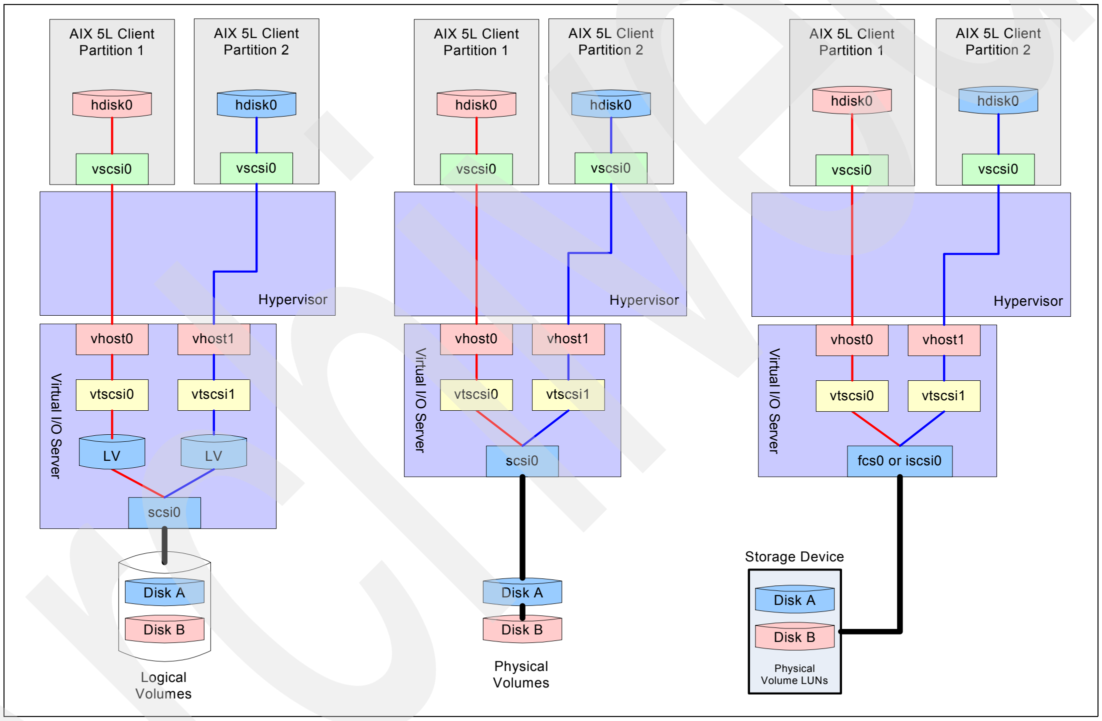


Figure 2-10 One Virtual I/O Server alternatives for backing storage

Description of scenario

In this scenario, a single Virtual I/O Server provides virtual disks to multiple client partitions. For example, the Virtual I/O Server can provide a virtual disk for each AIX 5L client partition's rootvg volume group. The Virtual I/O Server can provide as many disks to each client partition as required by that partition.

One of the advantages of using the Virtual I/O Server is that it provides a way to share a single physical disk with multiple clients. It is not a requirement to have an entire physical disk dedicated to each client partition, thus the Virtual I/O Server gives the ability to create multiple (boot or data) disks from a single physical disk.

The Virtual I/O Server has several options for providing the backing storage device for the virtual disks. In Figure 2-10, the backing devices for the virtual disks are logical volumes, physical volumes, and SAN LUN devices. These choices can be mixed and matched. The Virtual I/O Server is not limited to using just one type of backing storage device.

If using logical volumes for the backing storage device, on the Virtual I/O Server, a volume group (or storage pool in the case of Integrated Virtualization Manager systems) is created from one or more available physical volumes. Logical volumes are created on the volume group. Those logical volumes get mapped to the client partitions as virtual disks, one logical volume for each virtual disk. When creating the logical volume, the size specified for the logical volume will become the size of the virtual disk. This gives the flexibility to effectively use a few disk drives to support multiple partitions. Physical volumes are preferred when space allows and ultimate performance is a priority.

If using a whole physical disk for a backing device for your virtual disks, it is necessary to use a disk that the Virtual I/O Server has not assigned to a volume group. This approach presents the entire physical disk to the client partition as a virtual disk. This approach allows client partitions to share the physical storage controller but to use full performance of a physical disk. By using this approach, a failure of a single disk will only impact the single client partition using that disk.

If using SAN storage, the Virtual I/O Server can use a LUN for the storage backing device. The SAN environment can be used to provision and allocate LUNs and the Virtual I/O Server can map the LUN to a virtual disk as though it were a physical disk. If the LUN is configured with RAID, the disk being presented to the client partition is RAID protected by the SAN storage. If the client partitions require data protection, using a LUN configured as a RAID array can be an effective solution. It is preferred to map SAN-attached LUNs as physical volume backing storage rather then configuring volume groups and logical volumes on top of them.

If using the Integrated Virtualization Manager instead of the Hardware Management Console, the concept of logical volumes is replaced with storage pools. Like logical volumes, the storage pools can subdivide physical disks and present them to different clients as virtual disks. This approach can be used in entry-level server environments.

The alternatives for a backing storage device shown in this scenario indicate the flexibility of the Advanced POWER Virtualization feature when architecting the Virtual I/O Server virtual storage. Deciding which alternative is best depends on available disks and requirements from the client partitions.

> **Note:** This scenario uses a single path to the backing storage device. For a higher degree of data access or resilience, consider a scenario that uses multipath I/O or include a second Virtual I/O Server.

Figure 2-10 on page 48 illustrates the three backing storage alternatives:

- Logical volumes on a volume group
- Physical volumes on a single storage controller
- SAN storage LUN device

Architectural attributes

Table 2-8 describes the architectural attributes for each backing storage alternative. Remember it is possible for the Virtual I/O Server to use a combination of these alternatives.

Table 2-8 One Virtual I/O Server with single path to storage

	Disk access	
Resilience	Standard	🟩
Serviceability	Standard	🟩
Scalability/throughput	Standard	🟩
Flexibility	High	🟩🟩🟩

The following points explain the attributes and their rankings:

- Resilience

  Standard
  The use of logical volumes, physical volumes, or LUNs provides several options for presenting virtual disks to client partitions, but this configuration does not provide protection from a disk failure that might impact client partitions.

- Serviceability

  Standard
  Client partitions using a single path to their storage are impacted if the Virtual I/O Server is taken offline for maintenance.

- Scalability/throughput

  Standard
  Additional virtual disks can easily be added to support new client partitions as long as the physical disks or LUNs are available. The throughput of this configuration is inherent in the capability of the storage solution that is implemented.

- Flexibility

  High
  The creation of backing storage for virtual disks is highly flexible. The additional virtual disks can be mapped to new or existing client partitions.

Summary

The Virtual I/O Server has a lot of flexibility as seen by the choices you can make for a backing store device to mapped to a virtual disk. The decision of which alternative is best suited for a particular architecture is a function of the requirements of the client partition. Using logical volumes enables you to use a single physical disk for multiple client partitions. Using a physical volume gives you physical isolation of different client partitions. Using the SAN storage solutions gives you the flexibility and data protection that comes with the SAN implementation.

This configuration is suitable for solutions with the following characteristics:

- Client partitions with light to medium disk requirements.
- Consolidation of multiple small servers with small disk requirements on a single system.
- Dynamic changes in disk requirements to support new or existing client partitions.
- Storage efficiency is important.
- Dual Virtual I/O Servers are not affordable.

- Test, development environment, non-critical production systems.

Solutions with the following characteristics are not recommended:

- Access to data is required 24x7.
- Servers with high disk I/O requirements.

Further reference

For background and explanatory reading about the technologies mentioned in this scenario, see the following references:

- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768

2.2.3 One Virtual I/O Server with MPIO access to storage: Scenario 8

Figure 2-11 outlines the Advanced POWER Virtualization architecture using a single Virtual I/O Server and multiple path access to storage.

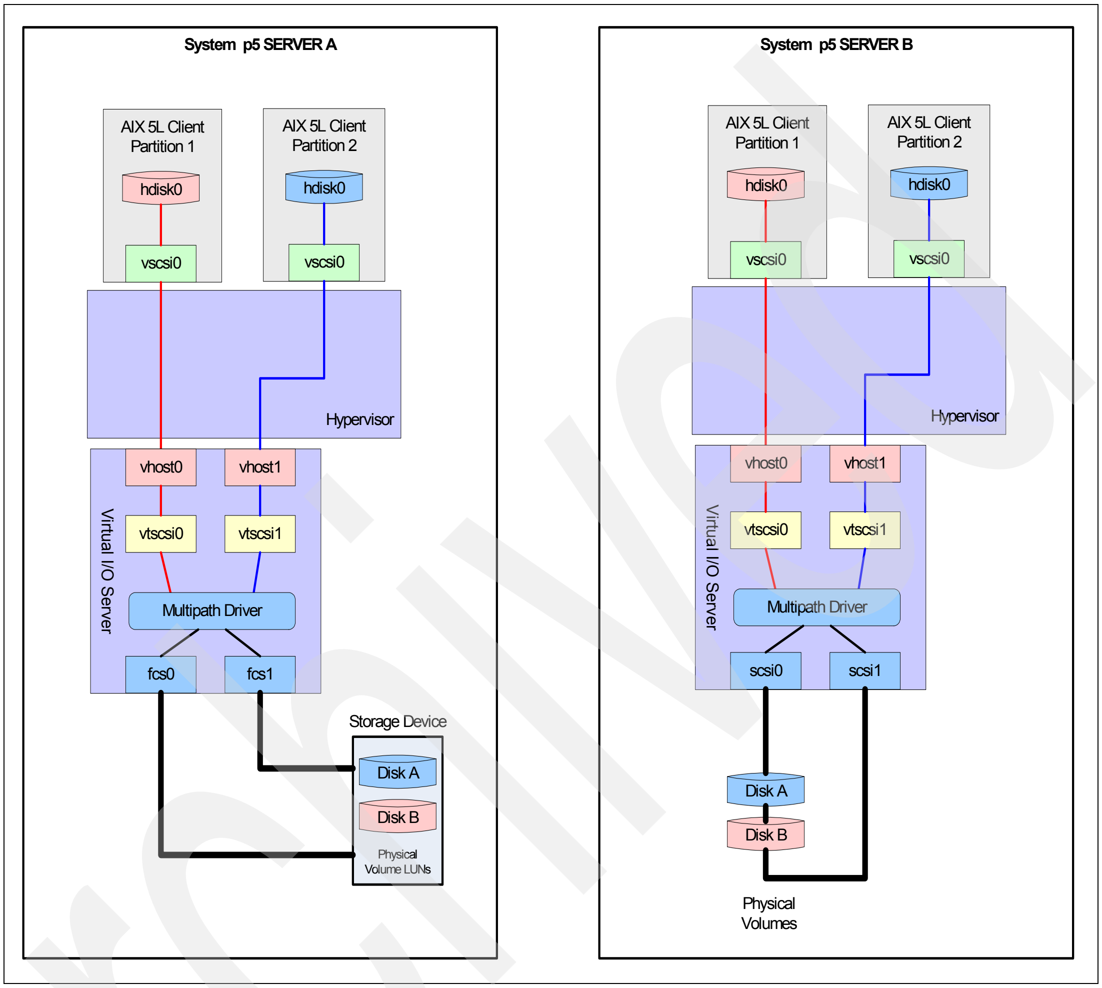


Figure 2-11 One Virtual I/O Server with MPIO access to storage

Description of the scenario

In this example scenario, one Virtual I/O Server provides access to storage devices for client partitions. A client partition uses a virtual SCSI client adapter paired with a virtual SCSI server adapter on the Virtual I/O Server side to access the data. Multiple physical paths to the storage device are provided for increased resiliency and throughput. This configuration provides increased resiliency against SAN path failure. That is, it can sustain a storage adapter failure and (in the appropriate SAN layout) a SAN switch failure.

The storage accessed can SCSI (server A in Figure 2-11) or SAN attached (server B in Figure 2-11). SCSI technology allows multipath access to disks (multiple adapters on a single SCSI bus), as shown on the diagram, but only fail-over mode is supported. With SAN technology, there is usually the possibility to use several paths to achieve required throughput with load balancing.

An appropriate multipath device driver for the given attached storage device must be installed on the Virtual I/O Server. We recommend using the standard MPIO driver if it is supported for the attached device (for example, SDDPCM is preferred over SDD for DS8xxx storage servers). A best practice is not to operate two distinct multipath drivers on the same fibre channel (FC) adapter. Therefore, if devices of various types are attached to a Virtual I/O Server, they must be accessed through distinct FC adapters.

Both logical volumes (LVs) and whole physical volumes (PVs) can be used as backing devices. In case of FC-attached devices, we recommend the use of whole LUNs. PVs allow a second Virtual I/O Server to access the same LUN.

Using this configuration, if the Virtual I/O Server is shut down for maintenance, access to storage from client partitions will not be possible. Resiliency and serviceability can be extended by adding a second Virtual I/O Server (2.2.5 “Dual Virtual I/O Servers with MPIO and client MPIO: Scenario 10” on page 60).

Architectural attributes

Table 2-9 provides the architectural attributes.

Table 2-9 One Virtual I/O Server with MPIO access to storage non-functional features

	Disk access	
Resilience	Medium	🟩🟩
Serviceability	Standard	🟩
Scalability/throughput	High	🟩🟩🟩
Flexibility	High	🟩🟩🟩

The following points explain the attributes and their rankings:

- Resilience

  Medium
  It can sustain a failure of a storage adapter or a SAN path, but if the Virtual I/O Server is shut down for maintenance or due to any unplanned outage, access to storage from the client partitions will not be possible.

- Serviceability

  Standard
  SAN maintenance can be performed as long as there is at least one path available, but if the Virtual I/O Server is taken offline for maintenance, access to storage from the client partitions will not be possible.

- Scalability/throughput

  High
  In a SAN environment, there are multiple paths in load balancing mode that are used to increase the total aggregated throughput. A number of these paths can be upgraded if more performance is needed.

  The exception is when using the RDAC driver and SCSI-attached devices. In that case, there is only one path active at any moment and thus throughput is limited to the throughput of a single adapter.

- Flexibility

  High
  Client partitions can be easily added without any additional hardware.

Summary, applicable workloads

This is a simple, highly flexible scenario with multipath access to storage that also provides increased resiliency against SAN path failures. If a load balancing algorithm is available in the multipath driver for a given storage device, the scenario is suited for very high throughput. Redundancy for the Virtual I/O Server is not provided, but if required, a secondary Virtual I/O Server can be implemented, as shown in 2.2.5 “Dual Virtual I/O Servers with MPIO and client MPIO: Scenario 10” on page 60.

This scenario can match many workloads where resilient access to storage and high throughput is required. The solutions using the scenario can range from low throughput to very high throughput and demanding applications.

This configuration is suitable for solutions with the following characteristics:

- There are no requirements about high resiliency and serviceability.
- Flexibility to add servers without additional disk attachment.
- Storage efficiency is important.
- Dual Virtual I/O Servers are not desired.
- Test, development, and production environments.
- Web, application, and database servers.

Solutions with the following characteristics are not recommended:

- Extremely high I/O throughput requirements
- High demand for resiliency and serviceability

Further reference

Advanced POWER Virtualization on IBM System p5, SG24-7940, describes supported configurations with MPIO.

2.2.4 Dual Virtual I/O Servers with AIX 5L client mirroring: Scenario 9

Figure 2-12 outlines the Advanced POWER Virtualization architecture using dual Virtual I/O Server and a single path to storage.

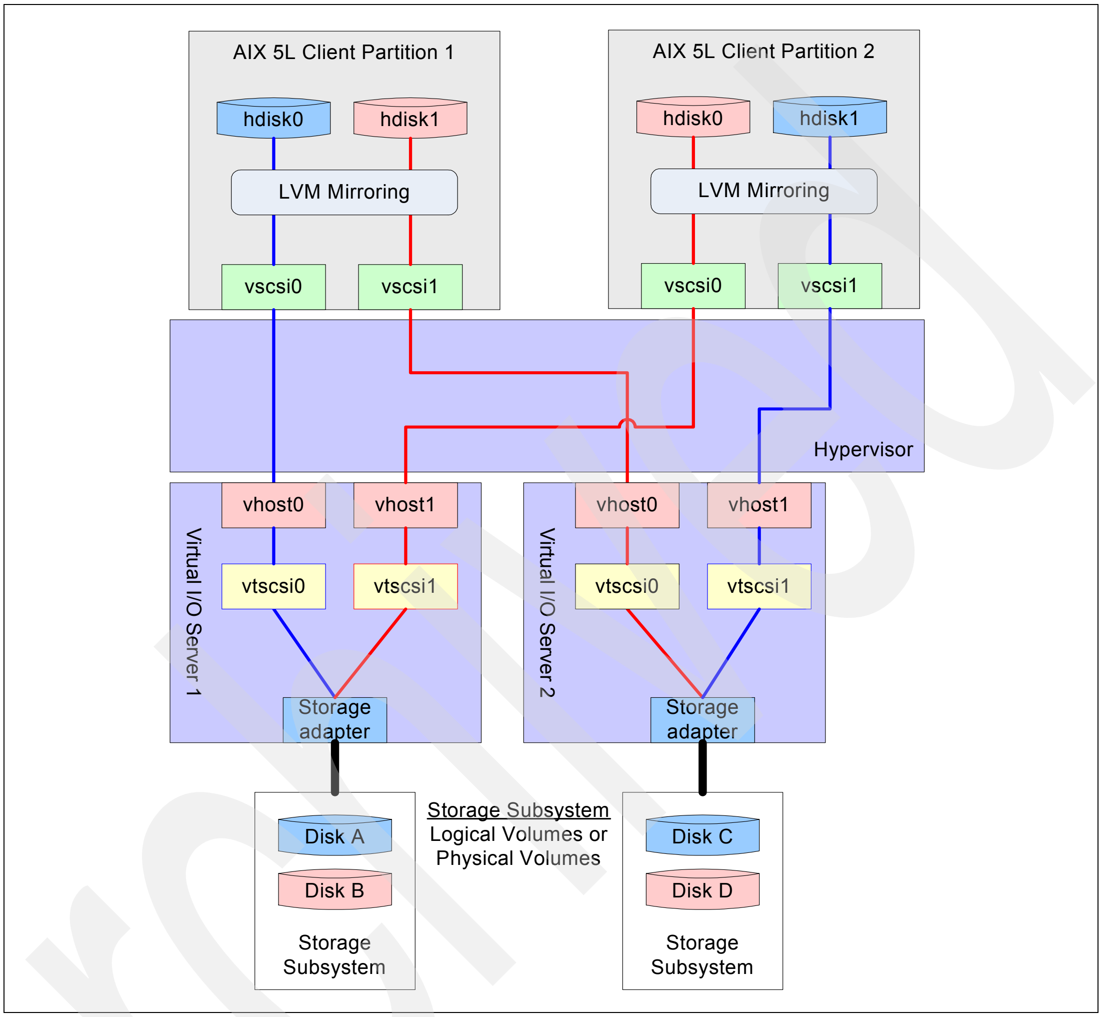


Figure 2-12 AIX client mirroring with dual Virtual I/O Servers, single physical path to storage

Figure 2-13 outlines the Advanced POWER Virtualization architecture using dual Virtual I/O Servers and multiple paths to storage.

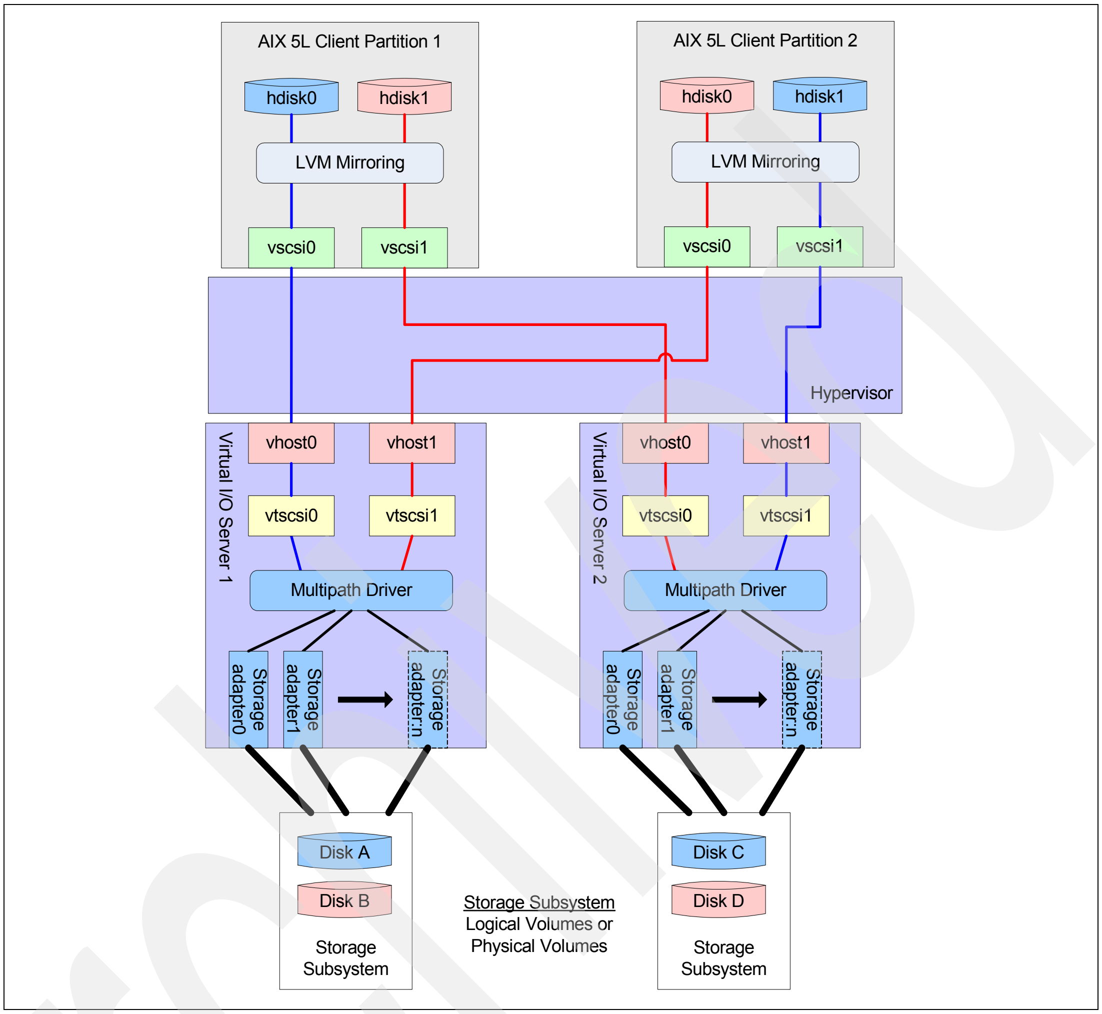


Figure 2-13 AIX client mirroring with dual Virtual I/O Servers, multiple physical paths to storage

Description of the scenario

In this example scenario, two Virtual I/O Servers host the storage devices for client partitions. The client partition uses Logical Volume Manager (LVM) mirroring to mirror its data between the disk sets. Each Virtual I/O Server presents to the client a virtual SCSI device that is connected to a different physical disk. Therefore, physical disks, storage adapters, and Virtual I/O Server redundancy is provided.

The physical storage can internal SCSI, external SCSI, or FC SAN-attached storage. Both LVs and whole PVs can be used as backing devices (in the case of FC-attached devices, we recommend the use of whole LUNs). Physical access from Virtual I/O Servers to storage can be either single path or multipath (MPIO, SDD, or SDDPCM, for example). We recommend using a multipath driver even if only a single physical path is implemented initially to ease later updates. You might later need to back up and restore the data in order to change from a single path solution to a multiple path solution because of incompatible disk identifiers.

Using this configuration, if a Virtual I/O Server is shut down for maintenance, a mirror is temporarily lost, but the client partition remains operating without any interruption. When the

Virtual I/O Server is restored, a manual resynchronization of the mirrors on the client partition must be performed.

The client partition is configured with two virtual SCSI client adapters, each of them paired with one virtual SCSI server adapter from each Virtual I/O Server. Virtual I/O Servers provide physical access to storage; the two storage devices (or physical disks in case of SCSI disks) are separate and cannot be shared between Virtual I/O Servers (although they can potentially be two different LUNs from the same SAN storage array).

> **Note:** There are configuration limitations when using LVM mirroring with two Virtual I/O Servers:
>
> - Mirror Write Consistency must be turned off.
> - Bad block relocation turned off.
> - No striping.

We show and evaluate two variants of the basic scenario: single path physical access from Virtual I/O Servers to storage and multipath access to storage.

Architectural attributes

In this scenario, we provide two diagrams. They present basically the same functionality, but they have different ratings in the architectural attributes tables (the multipath solution is more scalable and provides higher throughput). The tables and their text descriptions for individual diagrams follow.

AIX 5L client mirroring with dual Virtual I/O Servers, single path to storage

Table 2-10 provides the architectural attributes.

Table 2-10 Architectural attributes for AIX 5L client mirroring with dual VIOS, single path to storage

	Disk access	
Resilience	High	🟩🟩🟩
Serviceability	High	🟩🟩🟩
Scalability/throughput	Standard	🟩
Flexibility	High	🟩🟩🟩

The following points explain the attributes and their rankings:

- Resilience

  High
  It can survive the failure of any single infrastructure component.

- Serviceability

  High
  One Virtual I/O Server can be taken offline for maintenance without any outage to client partitions. Manual data resynchronization is then needed.

- Scalability/throughput

  Standard
  All client partitions share the throughput of one physical adapter in the virtual I/O client. Partitions can be added up to the point where the total aggregated throughput of a single SCSI or FC adapter is exceeded.

- Flexibility

  High
  Client partitions can be easily added without any additional hardware.

AIX 5L client mirroring with dual Virtual I/O Servers, multiple paths to storage

Table 2-11 provides the architectural attributes.

Table 2-11 Architectural attributes for AIX 5L client mirroring with dual VIOS, multiple paths to storage

	Disk access	
Resilience	High	🟩
Serviceability	High	🟩
Scalability/throughput	High	🟩
Flexibility	High	🟩

The following points explain the attributes and their rankings:

- Resilience

  High
  It can survive the failure of any single infrastructure component and even the failure of multiple storage attachment paths.

- Serviceability/Throughput

  High
  One Virtual I/O Server can be taken offline for maintenance without any outage to client partitions; manual data resynchronization is then needed.

- Scalability

  High
  Client partitions can be added easily. When total aggregated throughput requested by client partitions exceeds the given hardware configuration, additional storage adapters can be added.

  Requested throughput can be reached by placing enough physical adapters in the Virtual I/O Servers and by giving enough CPU entitlement to the Virtual I/O Server partitions.

- Flexibility

  High
  Client partitions can be added without any additional hardware.

Summary and applicable workloads

A highly available solution (in terms of disk access) that when combined together with one of the highly available network solutions (2.1.4 “One Virtual I/O Server with client-partition backup link: Scenario 3” on page 26, 2.1.6 “Dual Virtual I/O Servers with client LA backup channel: Scenario 5” on page 35, or 2.1.7 “Dual Virtual I/O Servers with SEA failover: Scenario 6” on page 40) provides as much resiliency as one physical IBM System p5 machine can provide. Individual Virtual I/O Servers can be brought down for maintenance without interrupting client partitions. Client partitions can be very easily added without any hardware change (if enough physical disk space is provided and if throughput of the actual number of disk storage physical adapters in the Virtual I/O Servers allows). We recommend using the multipath I/O driver for physical storage access even in the case of single path access because it facilitates easier an upgrade to a multipath environment. The scenario requires the allocation of twice the usable disk space required by client partitions. In the case of SAN-attached disk storage that has some level of redundancy already implemented, disk space usage might be less efficient.

This scenario can match many workloads where highly available access to storage and concurrent maintenance of Virtual I/O Server is required. The solutions using the scenario can range from low throughput to very high throughput and demanding applications.

This configuration is suitable for solutions with the following characteristics:

- High I/O throughput requirements (multipath diagram).
- Resiliency and serviceability on disk I/O requirements.
- There is no restriction about the storage capacity.
- Flexibility requirements where client partitions can be easily added without additional hardware change.
- Test, development, and production environment.
- Web, application, and database servers.

Solutions with the following characteristic are not recommended:

- Extremely high I/O throughput and low latency requirements

Further reference

For background and explanatory reading about the technologies mentioned in this scenario, see the following references:

- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940

2.2.5 Dual Virtual I/O Servers with MPIO and client MPIO: Scenario 10

Figure 2-14 depicts the disk configuration for this scenario.

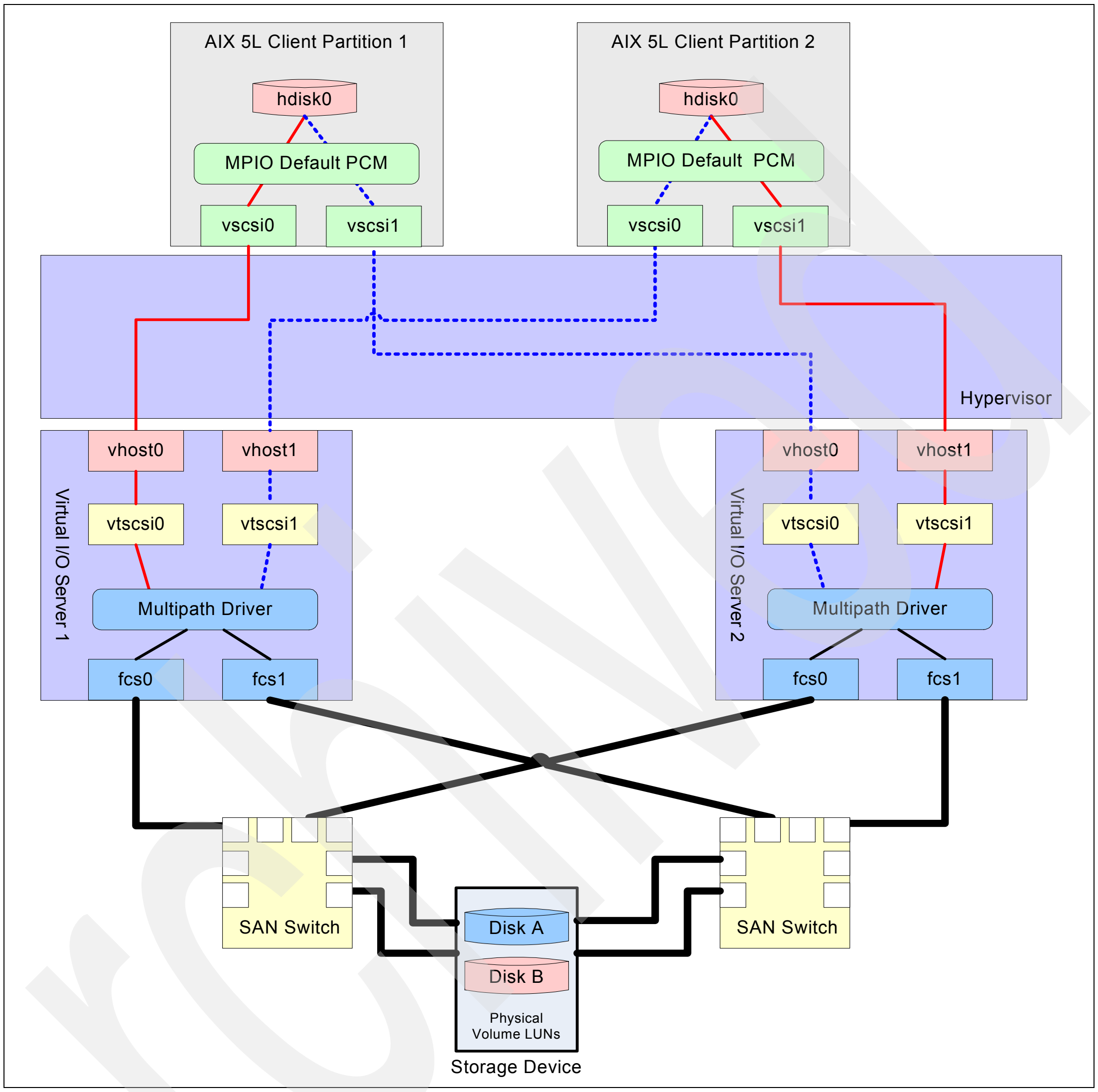


Figure 2-14 Dual Virtual I/O Servers using MPIO

Description of scenario

In this scenario, dual Virtual I/O Servers provided SAN-backed virtual disks to multiple client partitions. The configuration leverages multipath I/O (MPIO) to provide multiple paths to the data at the client partition as well as at the Virtual I/O Server. In addition, two SAN switches are included in the configuration to provide resilience in the SAN. This configuration provides multiple paths to the client data to help ensure data access under a variety of conditions. Although this configuration illustrates the virtualization of a single client disk, it can be extended to multiple disks if required.

Each Virtual I/O Server uses multiple paths to the SAN LUNs through two host bus adapters, each to different SAN switches. This allows the Virtual I/O Server to reach the data if either SAN switch fails. The SAN LUNs are treated as physical volumes and are mapped to two different virtual SCSI adapters to support two different client partitions. Each client partition has two virtual SCSI adapters. Each client virtual SCSI adapter connects to a different Virtual I/O Server. The client partition uses MPIO to provide separate paths to the data disk through each Virtual I/O Server. This way, the client partition can still get to its data if one of the Virtual I/O Servers is not available.

This configuration is very robust, but is more complicated than earlier configurations. It is necessary to set up MPIO on the client. Two Virtual I/O Servers need to be installed and configured. Two or more host bus adapters are needed in each Virtual I/O Server partition. The Virtual I/O Server is also configured for multipath I/O. The SAN LUNs are mapped as physical volumes, not logical volumes.

The architecture provides protection from various component outages whether they are planned (such as for maintenance) or unplanned. For example, it is possible to take one Virtual I/O Server offline for a code update without interrupting the client partitions. A SAN accessibility problem in one of the Virtual I/O Servers would not impact the client partitions. Even a SAN switch failure would not prevent the client partitions from getting to their data. Finally, by using RAID in the SAN LUNs, there is resilience to a disk failure.

With this architecture, there is also the potential for increased throughput due to multipath I/O. The primary LUNs can be split across multiple Virtual I/O Servers to help balance the I/O load.

Architectural attributes

Table 2-12 provides the architectural attributes.

Table 2-12 Dual Virtual I/O Servers using MPIO

	Disk access
Resilience	High 🟩
Serviceability	High 🟩
Scalability/throughput	High 🟩
Flexibility	High 🟩

The following points explain the attributes and their rankings:

- Resilience

  High
  This configuration provides a high degree of availability resilience of the architecture. By using two Virtual I/O Servers, there is protection if a Virtual I/O Server is taken offline. Using two host bus adapters provides protection if one of the adapters fails. The two SAN switches provide protection from a switch or data path failure. Finally, using RAID-protected disks protects against a single physical disk failure.

- Serviceability

  High
  The configuration is highly serviceable for the same reasons it is highly available. The redundancy in this design allows the client partitions to continue to run while individual components in this configuration are serviced.

- Scalability/throughput

  High
  This configuration provides a high degree of scalability with its ability to dynamically support additional client partitions. The Virtual I/O Servers can add new virtual SCSI adapters that can be mapped to new clients. There is also the ability to increase I/O throughput due to multipath I/O. Primary LUNs can be split across multiple Virtual I/O Servers to help balance the I/O workload.

- Flexibility

  High
  This configuration is highly flexible because, without any additional hardware, SAN LUNs can be configured to support new client partitions without disrupting the existing client partitions.

Summary

The use of dual Virtual I/O Servers, along with multiple adapter cards and multipath access to data, yield a very robust solution for client partitions requiring high levels of data access. This configuration is one example of how the Advanced POWER Virtualization technology can be leveraged to satisfy this need while making effective use of the hardware.

Combining this storage configuration with one of the highly available network solutions, such are the one described in 2.1.7 “Dual Virtual I/O Servers with SEA failover: Scenario 6” on page 40, yields a complete, robust network and storage solution.

This scenario can match many workloads where highly available access to storage and concurrent maintenance of Virtual I/O Server is required. The solutions using the scenario can range from low throughput to very high throughput and demanding applications.

This configuration is suitable for solutions with the following characteristics:

- There are requirements for high I/O throughput, serviceability, and resilience.
- There are no hardware resource restrictions.
- Flexibility requirements in both adding additional servers and extending storage capacity.
- Dual Virtual I/O Servers are desired.
- Test, development, and production environments.
- Web, application, and database servers.

Solutions with the following characteristic are not recommended:

- Extremely high I/O throughput requirements

We recommend using Virtual I/O Server Version 1.3 when high virtual SCSI I/O throughput is needed to take advantage of the performance improvements included in that version of the server.

Further reference

For background and explanatory reading about technologies mentioned in this scenario, see the following references:

- “SAN environment LUN identification” on page 47
- Supported Virtual I/O Server storage subsystem solutions, available at:

  `http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/VIOS_datasheet_081706.html`
- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940

2.2.6 Dual Virtual I/O Servers with MPIO, client MPIO, and HACMP: Scenario 11

Figure 2-15 outlines the Advanced POWER Virtualization architecture together with High-Availability Cluster Multi-Processing (HACMP) architecture using double Virtual I/O Servers.

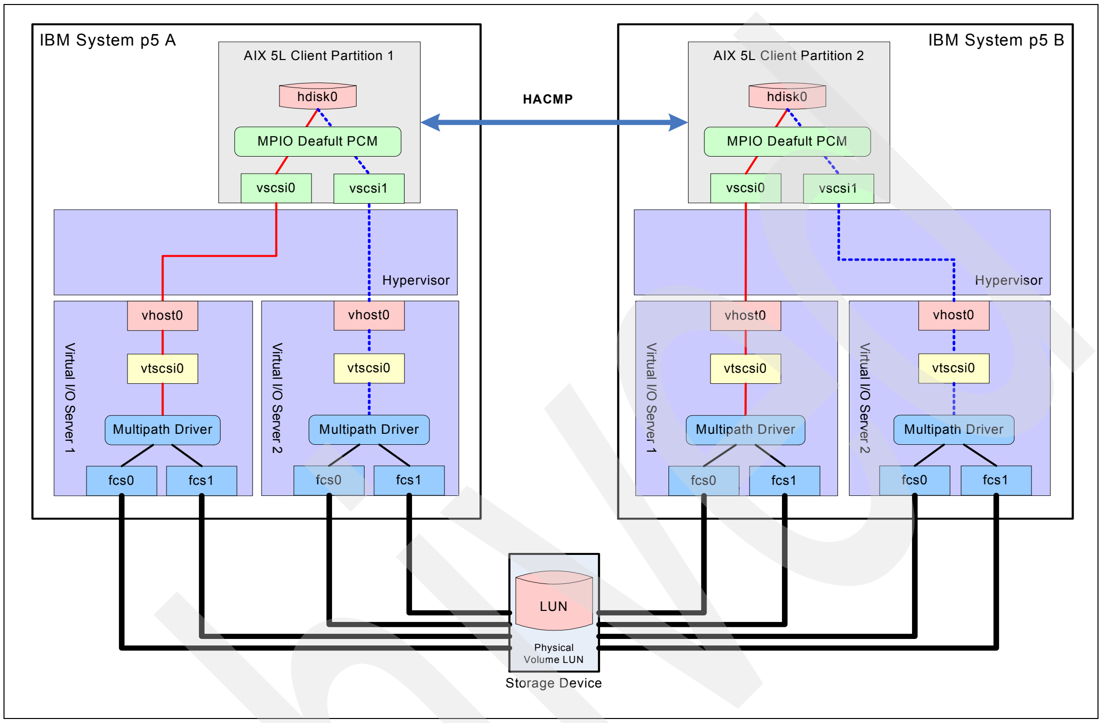


Figure 2-15 AIX 5L client mirroring with HACMP and dual VIOS, multiple physical paths to storage

Description of the scenario

In this more complex scenario, we use a cluster of two client partitions in two separate IBM System p5 servers in order to achieve cluster-level resiliency and serviceability. This scenario extends scenario 10 (2.2.5 “Dual Virtual I/O Servers with MPIO and client MPIO: Scenario 10” on page 60) with an additional IBM System p5 server and HACMP software. Storage is shared and can be accessed from both client partitions (cluster nodes) either concurrently or non-concurrently.

In each IBM System p5 server, dual Virtual I/O Servers provide access to SAN storage. The configuration leverages MPIO to provide multiple paths to the data at the client partition as well as at the Virtual I/O Server. The SAN infrastructure is simplified for the purpose of the diagram. In actual implementations, at least two switches are included in the configuration to provide high resilience in a SAN environment.

In order to provide protection against data loss, this scenario often is extended with synchronous or asynchronous remote data mirroring over geographically dispersed locations. This is not included in the diagram and is often provided by storage hardware equipment (for example, the IBM Metro Mirror feature in the IBM DS8000™ storage server series).

A brief description of HACMP cluster software follows.

HACMP stands for High-Availability Cluster Multi-Processing and is a well-established IBM software product. It offers:

- Resiliency through automation, providing unattended recovery from failures

- Improved availability of applications, including health check of running applications
- A scalable growth strategy for shared workloads
- Auto-discovery and auto-management features that simplify implementation and management

HACMP middleware allows a group of cluster nodes (partitions) to cooperate and offer increased overall availability and transactional performance. HACMP exploits redundancy of all critical elements to ensure automatic recovery in the case of a single failure of any hardware or software component. It is implemented on each AIX 5L system. From a minimum of two systems, the cluster can be extended to 32 systems. In the basic two node configuration, one is the production node, the other is the backup node. If the production system becomes unavailable, the backup system automatically activates itself and takes over production system resources. HACMP also supports more advanced configurations, such as a “cascading configuration,” where more than two systems are clustered and each system runs different applications. If one system is failing, another restarts that application load on its own. HACMP can run in fail-over mode or in concurrent mode, where several systems access the shared space concurrently. For a more detailed description, see:

http://www.ibm.com/systems/p/software/hacmp.html

In this scenario (see Figure 2-15 on page 63), two Virtual I/O Servers in each IBM System p5 machine provide access to the storage devices for a client partition. The client partition is a member of an HACMP cluster. There is one pair of virtual SCSI client and virtual SCSI server adapters between each client partition and Virtual I/O Server. The client partition uses a standard MPIO driver with fail-over feature. The Virtual I/O Server uses MPIO (appropriate MPIO driver for given storage device) to load balance I/O between adapters.

The configuration provides protection for every single component outage (except the storage device itself) whether planned, such as for maintenance, or unplanned, such as a hardware failure. For example, it is possible to take one Virtual I/O Server offline for a code update without interrupting the client partitions. An adapter card failure in one of the Virtual I/O Servers does not impact the client partitions. Even a switch failure does not prevent the client partitions from getting to their data. If the SAN LUNs are configured as RAID arrays, even a disk failure will not cause an outage to the client partition. In addition, the whole IBM System p5 server can be shut down for maintenance if required and HACMP will handle it with its procedures; production will be taken over by another member of the HACMP cluster on another IBM System p5 machine.

With this scenario, there is the potential for very high bandwidth due to multipath I/O in the Virtual I/O Servers. In addition, the access to individual LUNs from the client partition can be split across the two Virtual I/O Servers to help balance the I/O load.

Architectural attributes

Table 2-13 provides the architectural attributes.

Table 2-13 Dual Virtual I/O Servers with MPIO, client MPIO, and HACMP

	Disk access	
Resilience	Clustered	🟩🟩🟩🟩
Serviceability	Clustered	🟩🟩🟩🟩
Scalability/throughput	High	🟩🟩🟩
Flexibility	High	🟩🟩🟩

The following points explain the attributes and their rankings:

- Resilience

  Cluster
  The scenario provides cluster-level protection from every single component outage except the storage device. Either of the two IBM System p5 servers can be shut down, and the production load will be taken over (manually or automatically) by the other active server.

- Serviceability

  Cluster
  The scenario provides cluster-level serviceability. Maintenance on every single component can be done concurrently with the production workload.

- Scalability/throughput

  High
  Client partitions can be added easily. When the total aggregated throughput requested by the client partitions exceeds given hardware configuration, additional storage adapters can be added.

  Required throughput can be reached by placing enough physical adapters to Virtual I/O Servers and by giving enough CPU entitlement to the Virtual I/O Server partitions.

- Flexibility

  High
  Client partitions can be added without additional hardware. The client partitions can be migrated between the two IBM System p5 servers.

Summary and applicable workloads

The use of cluster software together with dual Virtual I/O Servers and multiple adapter cards and multipath access to data brings the resiliency and serviceability of the whole solution to cluster level. The configuration provides protection from single component outages, except for the storage device, and yields a very robust solution for client partitions requiring high levels of data access. This configuration is one example how the Advanced POWER Virtualization technology can be leverage to satisfy very high application requirements while making effective use of the hardware.

This scenario can match many workloads where highly available access to storage and concurrent maintenance of Virtual I/O Server is required. The solutions using the scenario can range from low throughput to very high throughput and demanding applications.

This configuration is suitable for solutions with the following characteristics:

- There are requirements for high I/O throughput, serviceability, and resiliency.
- Cluster-level serviceability and resiliency with physical server redundancy requirements, protection for every single infrastructural component outages.
- Maintenance on every single component must be done concurrently with the production workload.
- There are no hardware resource restrictions such as systems, I/O adapters, and storage capacity.
- Dual Virtual I/O Servers are desired.
- Cluster test, development, and production environments.
- Application and database servers.

Solutions with the following characteristic are not recommended:

- Extremely high I/O throughput requirements

Further reference

For more detailed information about HACMP, refer to:

- HACMP for System p

  http://www.ibm.com/systems/p/software/hacmp.html
- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940

3

Deployment case studies

The Advanced POWER Virtualization technology has an extensive list of possible configurations. The combinations of virtual SCSI and VLAN has enabled many organizations, large and small, to achieve flexibility not seen before on UNIX-enabled hardware. The examples in this chapter demonstrate the flexibility of the Advanced POWER Virtualization technology and use many of the previously described scenarios. These examples are from IBM client testimonials or provided by existing clients representing real-world problems and in-production solutions.

> **Important:** At the time of writing this document, all the licensing configurations for the products mentioned were validated. However, check with your software vendor or representative to ensure that the same software licensing structure applies to your configuration.

3.1 Global banking institution uses Integrated Virtualization Manager on a JS21 blade

In this section, we discuss the challenges, solution, and benefits of this configuration.

3.1.1 Challenges

The client was running Oracle® and AppWorks (Open Systems Scheduling application) on two Linux Vmware Intel®-based systems each with two CPUs for development and testing environments. The production environment runs in existing IBM System p hardware that had issues due to development test systems running a different operating system compared to production. This was most noticeable through management processes such as upgrading and patching, and it created further issues with maintaining separate patch repositories for each operating system and application. The key points in migrating the development and test system to IBM System p hardware are:

- Keep hardware costs low.
- Reuse existing hardware.
- Reduce software licensing costs.
- Provide similar or better response.
- Provide a development and test environment similar to production in software revisions.

3.1.2 Solution

The client decided that it would be cost effective to use spare blade chassis capacity by adding an IBM BladeCenter® JS21 blade. The mix of Intel and pSeries® blades is supported by IBM and would not impact existing blades in the chassis. In addition, the Integrated Virtualization Manager was used to split the JS21 hardware into four separate partitions consisting of a Virtual I/O Server and three other partitions to be used by development and testing team.

3.1.3 Benefits

The client obtains the following benefits with this configuration:

- Development, test, and staging environments were still separated into partitions.
- Development and testing mirrored more closely the production software revisions.
- Patching and upgrading can now be accurately tested and documented, providing a process to apply in production.
- Performance was equivalent to running on separate systems.
- Oracle workgroup licenses were reduced.

3.1.4 Architectural design and diagrams

Figure 3-1 on page 69 shows the JS21 partition diagram, and Figure 3-2 on page 69 shows the hardware breakdown.

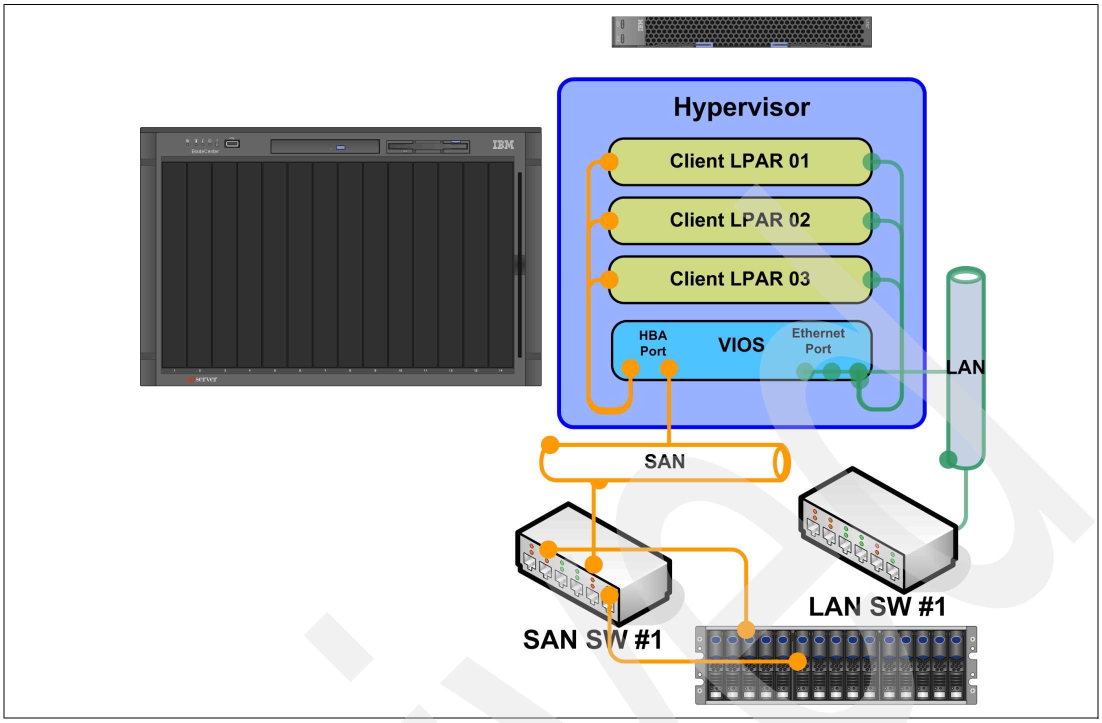


Figure 3-1 JS21 partition diagram

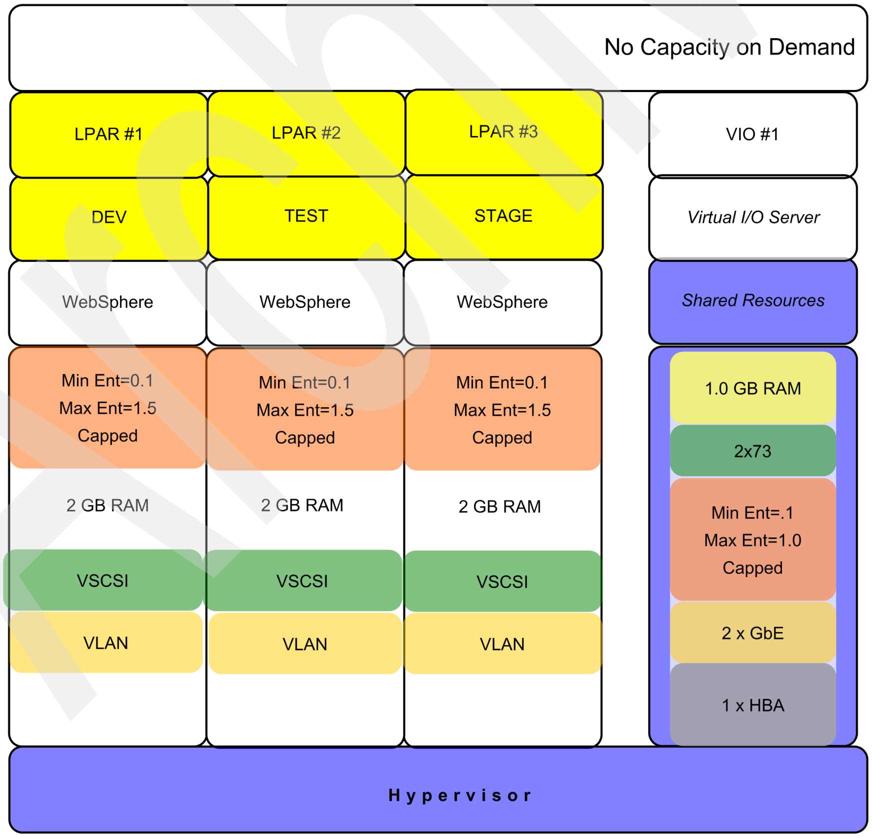


Figure 3-2 JS21 hardware breakdown

3.2 Leading Australian bank manages for growth on p5-510

In this section, we discuss the challenges, solution, and benefits of this configuration.

3.2.1 Challenges

The client found that the growing need to test and deploy new applications on IBM WebSphere® Application Server software placed pressure on the budget through software licensing and had an ever increasing demand on computer hardware, computer room space, power, and support staff. Normally to develop production applications, they used an extensive test and development regime that migrated each application through 10 separate environments, adding pressure on resources. The key points in this consolidation are:

- Consolidate disperse WebSphere Application Servers for development and testing into one system.
- Provide a separate physical operating system for each stage of application deployment.
- Allow scaling and sharing of CPU for each partition as needed or for short-term volume tests.
- Reduce licensing costs of WebSphere licenses.
- Prevent a performance loss.

3.2.2 Solution

Working with an IBM Business Partner, the client decided that rather than purchase 10 physical systems and pay licensing for 10 or more WebSphere licenses, it is more cost effective to purchase one quad-core IBM System p5 510 server and pay only four WebSphere licenses. Then, using Integrated Virtualization Manager split the system into 10 separate partitions sharing the resources.

3.2.3 Benefits

The client obtains the following benefits with this configuration:

- WebSphere licenses were reduced.
- Partitions share network resources, CPU, disk, and as a side benefit, can easily migrate between different test network zones without re-patching.
- The p5-510 used only 2U, reducing the footprint and power requirements.
- Performance is comparable to separate hardware installations.

3.2.4 Architectural design and diagrams

Figure 3-3 on page 71 shows the partition configuration, and Figure 3-4 on page 71 provides the hardware breakdown.

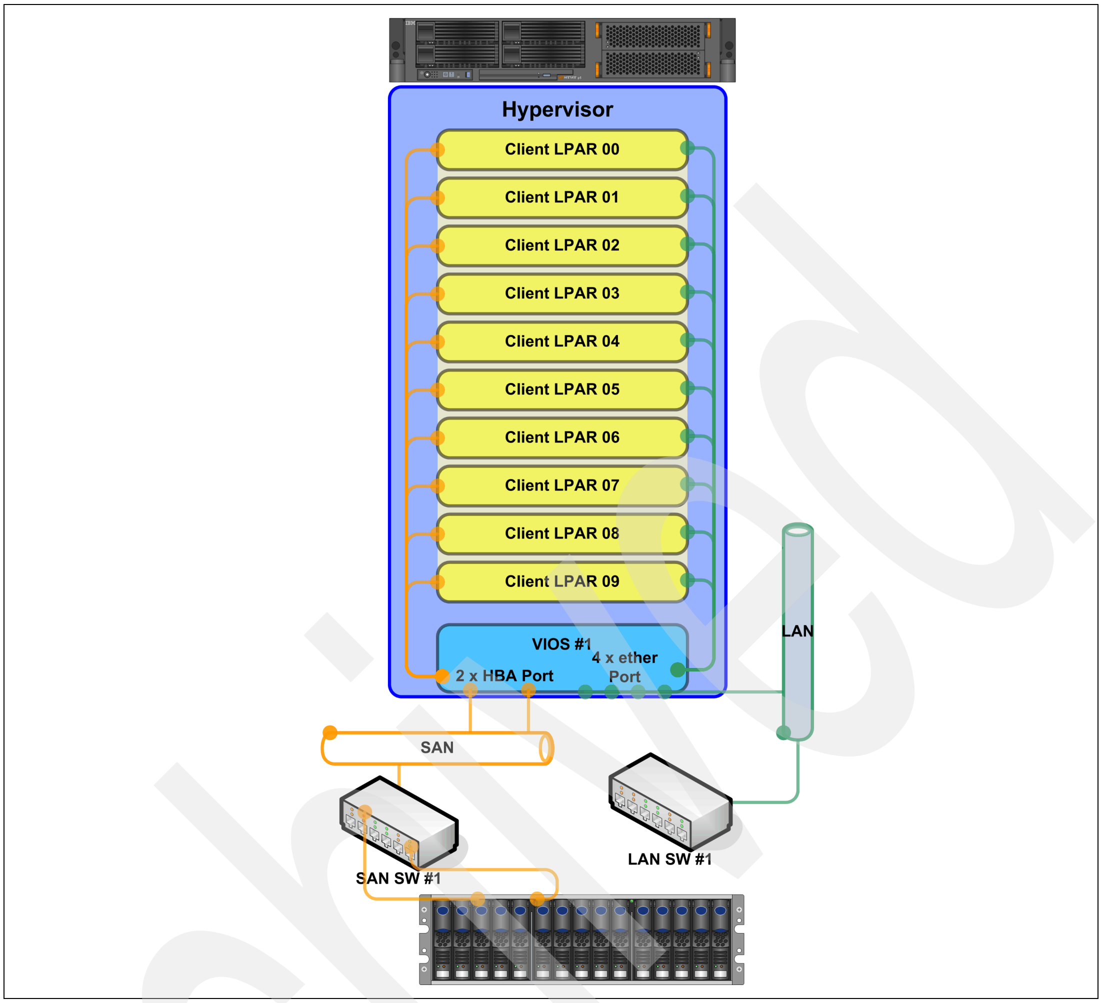
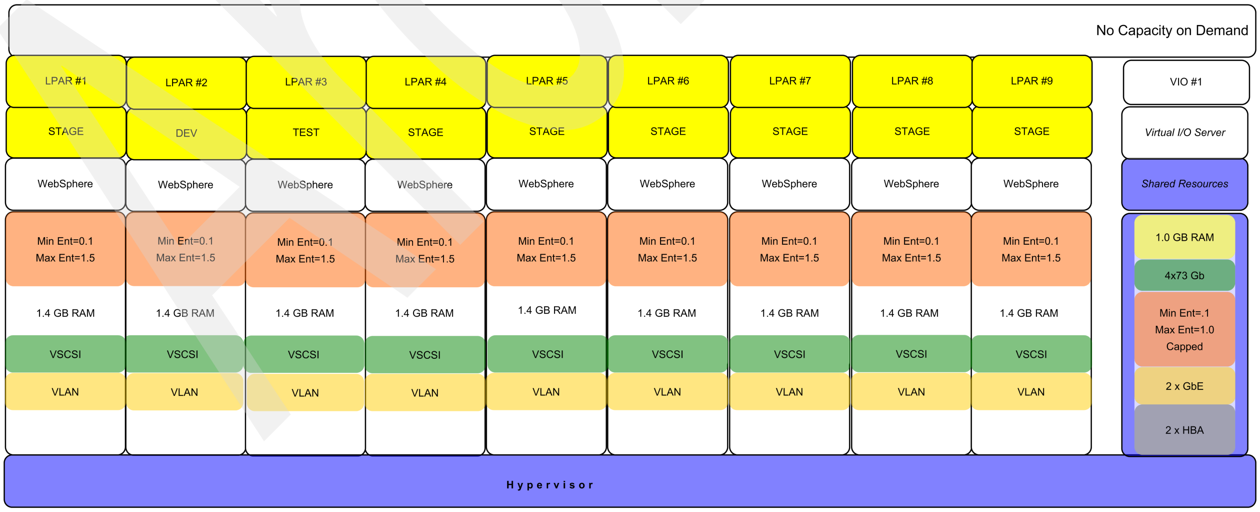


Figure 3-3 p5-510 split into 10 partitions sharing LAN and disk adapters



No Capacity on Demand

LPAR #1	LPAR #2	LPAR #3	LPAR #4	LPAR #5	LPAR #6	LPAR #7	LPAR #8	LPAR #9	VIO #1
STAGE	DEV	TEST	STAGE	STAGE	STAGE	STAGE	STAGE	STAGE	*Virtual I/O Server*
WebSphere	WebSphere	WebSphere	WebSphere	WebSphere	WebSphere	WebSphere	WebSphere	WebSphere	*Shared Resources*
Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	Min Ent=0.1 Max Ent=1.5	1.0 GB RAM 4x73 Gb
1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	1.4 GB RAM	Min Ent=.1 Max Ent=1.0 Capped
VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	
VLAN	VLAN	VLAN	VLAN	VLAN	VLAN	VLAN	VLAN	VLAN	2 x GbE
									2 x HBA

Hypervisor

Figure 3-4 p5-510 hardware breakdown

3.3 Financial sector combines DR and development systems

In this section, we discuss the challenges, solution, and benefits of this configuration.

3.3.1 Challenges

A leading Australian bank deploys PeopleSoft® as client relationship manager. To prepare for disaster recovery (DR), an equivalent system is at a remote site. To complicate the deployment, an extremely short time frame was established (six months) with the following requirements:

- Confirm sizing of hardware and convert to specification to order. Hardware must be able to serve greater than 5000 concurrent users and up to 7000 at peak loads.
- Ensure that expected response times are met.
- Provide dynamic environment to scale up and down partitions.
- Allow quick delivery of new test systems.
- DR hardware must be quickly and easily converted to act as production if needed.
- Engage as much of the spare or unused capacity as possible.

3.3.2 Solution

Working with an IBM Business Partner, the client purchased two IBM System p5 590 (p5-590) servers. To fully use all the hardware purchased, the client reused the disaster recovery hardware in a remote site for the development, testing, and deployment of the PeopleSoft application. Advanced POWER Virtualization is used on the disaster recovery hardware to create six separate client partitions, one used for each stage of deployment and three standby partitions to be used in case of disaster recovery.

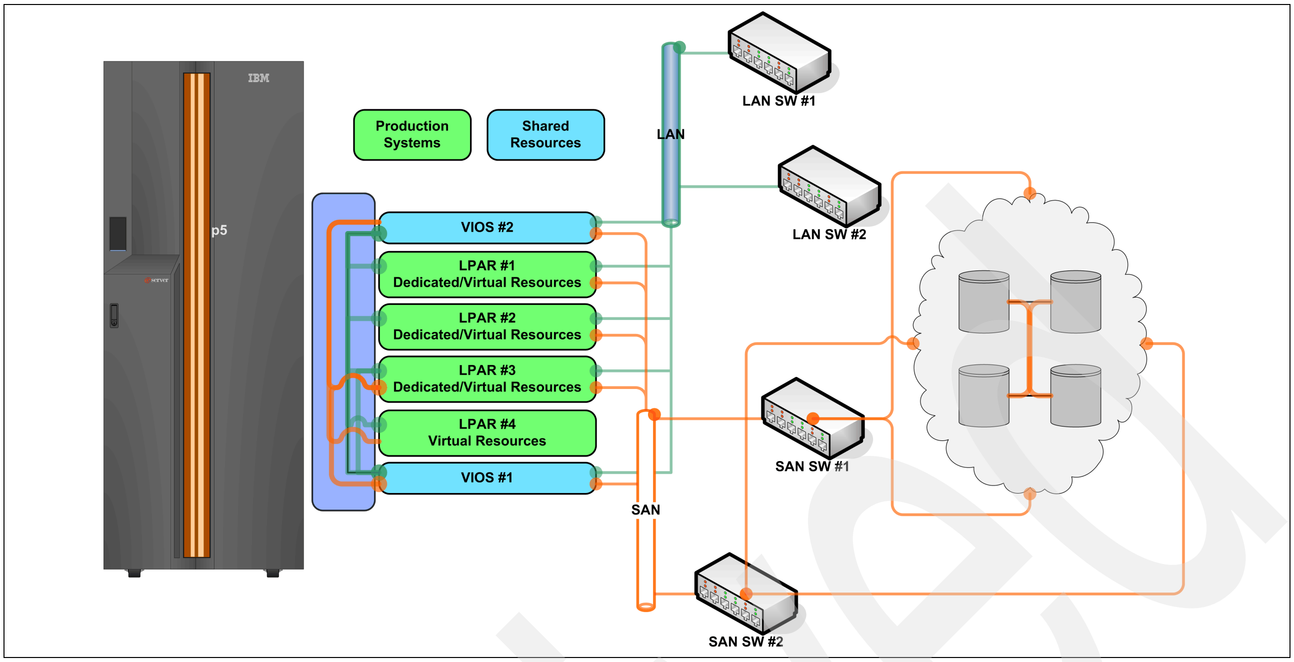
3.3.3 Benefits

The client was able to accurately estimate the hardware needed to handle the workload, and IBM guaranteed the response times on the IBM System p5 server. Standby capacity was provided to handle organic growth in the application functionality and, if needed, more applications due for rollout shortly after. The disaster recovery frame was fully used for the application deployment functions during the normal business hours. After hours, it was then used for stress testing new functionality by scaling down all non-essential partitions and scaling up stress, volume, and performance partitions.

Keeping the production hardware configuration separate from the development functions meant that mandatory disaster recovery tests were easily and quickly performed. During these tests, non-production partitions are shut down and all active capacity is used by disaster recovery partitions. IBM also offers the flexibility to enable all Capacity on Demand for short periods so that full production loads can be tested on the disaster recovery infrastructure.

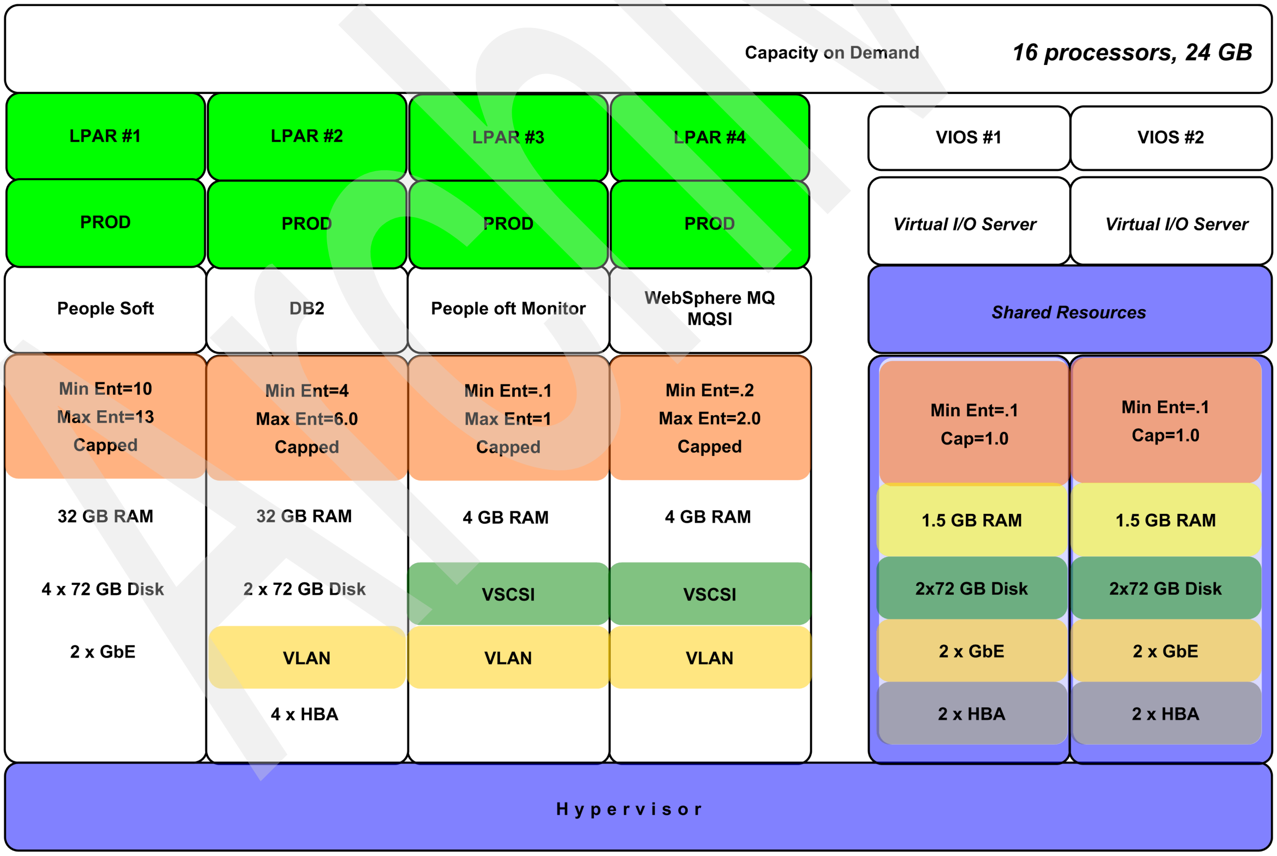
3.3.4 Architectural design and diagrams

Figure 3-5 on page 73 shows the partition configuration, and Figure 3-6 on page 73 provides the hardware breakdown. A second set of diagrams (Figure 3-7 on page 74 and Figure 3-8 on page 74) provide the recovery environments.



IBM
p5
Production Systems
Shared Resources
VIOS #2
LPAR #1 Dedicated/Virtual Resources
LPAR #2 Dedicated/Virtual Resources
LPAR #3 Dedicated/Virtual Resources
LPAR #4 Virtual Resources
VIOS #1
LAN
LAN SW #1
LAN SW #2
SAN SW #1
SAN
SAN SW #2

Figure 3-5 Production p5-590 with three partitions, all with mixed virtual and dedicated devices and CPU



Capacity on Demand				16 processors, 24 GB	
LPAR #1	LPAR #2	LPAR #3	LPAR #4	VIOS #1	VIOS #2
PROD	PROD	PROD	PROD	*Virtual I/O Server*	*Virtual I/O Server*
People Soft	DB2	People oft Monitor	WebSphere MQ MQSI	*Shared Resources*	
Min Ent=10 Max Ent=13 Capped	Min Ent=4 Max Ent=6.0 Capped	Min Ent=.1 Max Ent=1 Capped	Min Ent=.2 Max Ent=2.0 Capped	Min Ent=.1 Cap=1.0	Min Ent=.1 Cap=1.0
32 GB RAM	32 GB RAM	4 GB RAM	4 GB RAM	1.5 GB RAM	1.5 GB RAM
4 x 72 GB Disk	2 x 72 GB Disk	VSCSI	VSCSI	2x72 GB Disk	2x72 GB Disk
2 x GbE	VLAN	VLAN	VLAN	2 x GbE	2 x GbE
	4 x HBA			2 x HBA	2 x HBA
Hypervisor					

Figure 3-6 Hardware breakdown of production p5-590

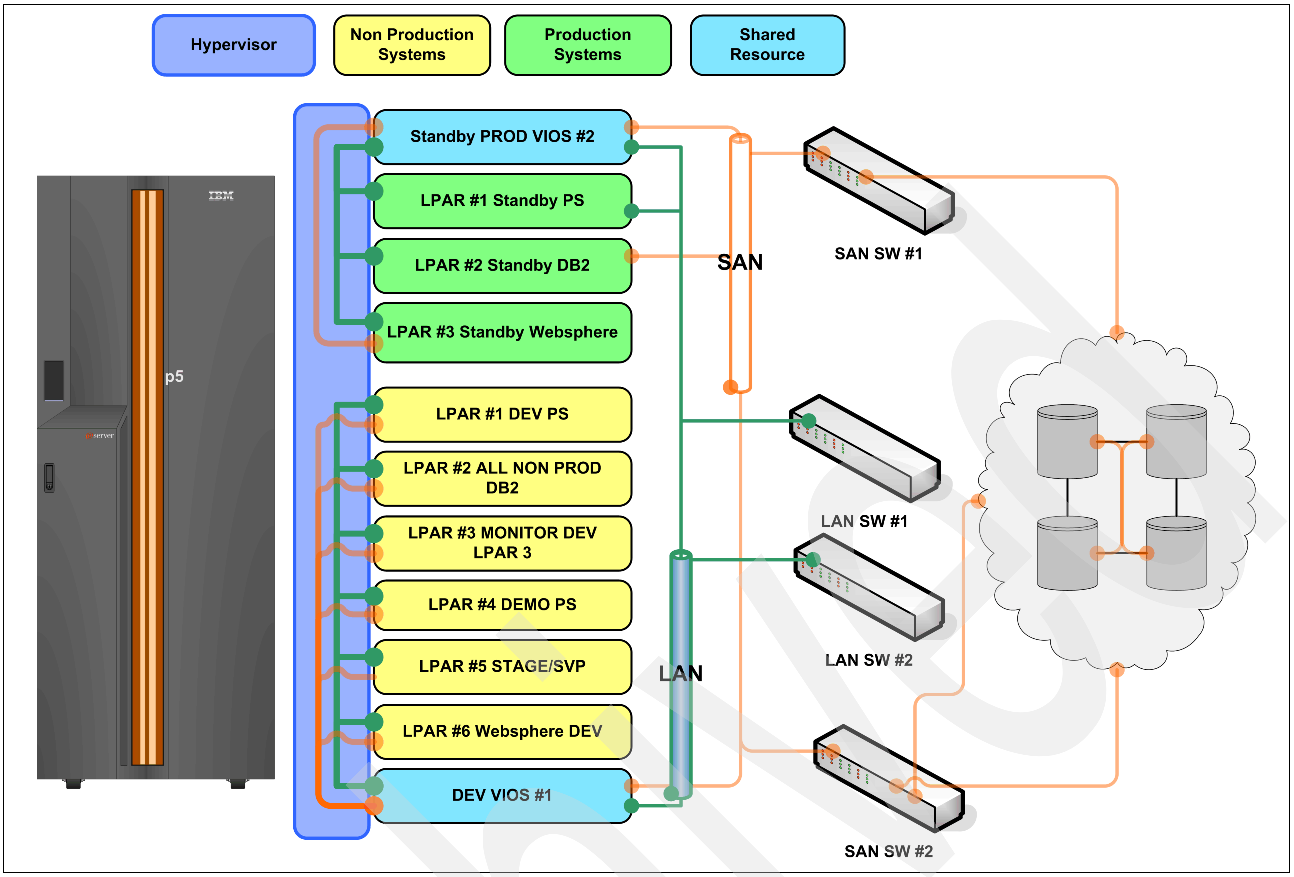
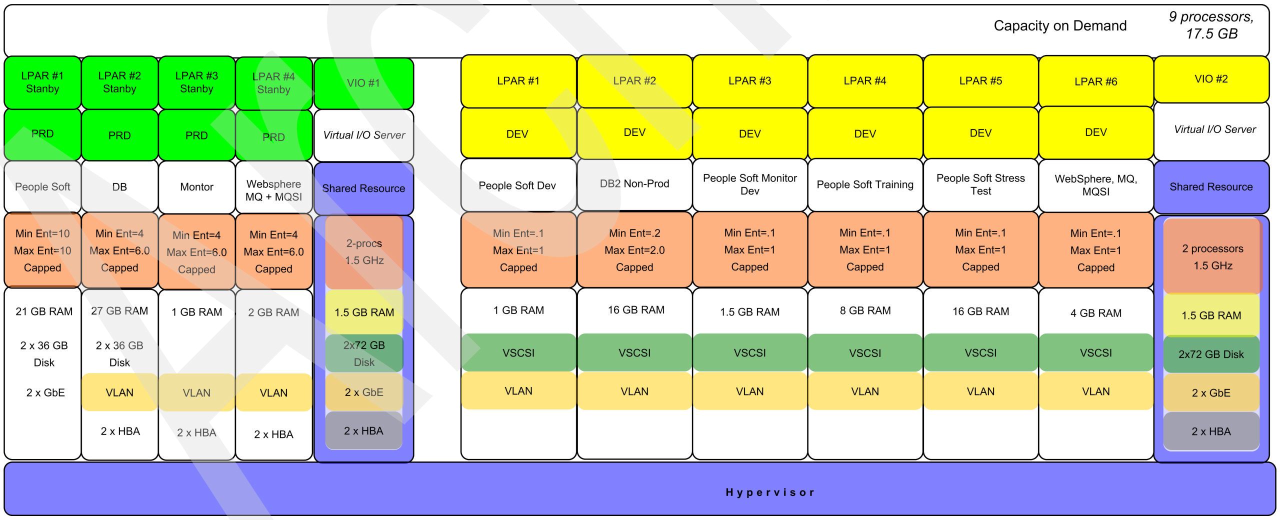


Figure 3-7 Second p5-590 for disaster recovery, test, and development



											Capacity on Demand	9 processors, 17.5 GB
LPAR #1 Stanby	LPAR #2 Stanby	LPAR #3 Stanby	LPAR #4 Stanby	VIO #1	LPAR #1	LPAR #2	LPAR #3	LPAR #4	LPAR #5	LPAR #6	VIO #2	
PRD	PRD	PRD	PRD	Virtual I/O Server	DEV	DEV	DEV	DEV	DEV	DEV	Virtual I/O Server	
People Soft	DB	Montor	Websphere MQ + MQSI	Shared Resource	People Soft Dev	DB2 Non-Prod	People Soft Monitor Dev	People Soft Training	People Soft Stress Test	WebSphere, MQ, MQSI	Shared Resource	
Min Ent=10 Max Ent=10 Capped	Min Ent=4 Max Ent=6.0 Capped	Min Ent=4 Max Ent=6.0 Capped	Min Ent=4 Max Ent=6.0 Capped	2-procs 1.5 GHz	Min Ent=.1 Max Ent=1 Capped	Min Ent=.2 Max Ent=2.0 Capped	Min Ent=.1 Max Ent=1 Capped	Min Ent=.1 Max Ent=1 Capped	Min Ent=.1 Max Ent=1 Capped	Min Ent=.1 Max Ent=1 Capped	2 processors 1.5 GHz	
21 GB RAM	27 GB RAM	1 GB RAM	2 GB RAM	1.5 GB RAM	1 GB RAM	16 GB RAM	1.5 GB RAM	8 GB RAM	16 GB RAM	4 GB RAM	1.5 GB RAM	
2 x 36 GB Disk	2 x 36 GB Disk			2x72 GB Disk	VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	VSCSI	2x72 GB Disk	
2 x GbE	VLAN	VLAN	VLAN	2 x GbE	VLAN	VLAN	VLAN	VLAN	VLAN	VLAN	2 x GbE	
	2 x HBA	2 x HBA	2 x HBA	2 x HBA							2 x HBA	
Hypervisor												

Figure 3-8 Hardware breakdown of recovery p5-590

3.4 Server consolidation reduces total cost of ownership

A large entertainment media company estimates that it will reduce its server costs by 20% through lower maintenance costs, decreased power and facilities expenses, and reduced software-licensing costs when it engages IBM Business Partner Direct Systems Support to replace its aging server platform with IBM eServer™ p5 570 servers running the IBM AIX 5L V5.2 and V5.3 operating platforms.

3.4.1 Challenges

Supporting its large-scale business operations, the client was relying on an IBM competitor's large server infrastructure. The three-year-old server platform was becoming dated and it could not support many of the features that come standard with newer server technologies. Furthermore, the systems required many central processing units (CPUs) that drove up software and maintenance costs, made the environment difficult to manage, caused high power and cooling costs, and occupied a very large footprint. By replacing its aging and expensive hardware with a new server infrastructure, the client expected to decrease the footprint of its server environment, lower its total cost of ownership (TCO), and improve its overall server performance.

3.4.2 Solution

The client engaged IBM Business Partner Direct Systems Support to design and deploy a new server infrastructure based on the IBM System p5 server platform. Direct Systems Support consolidated the client's 32 servers onto nine 16-way p5-570 servers running the IBM AIX 5L V5.2 and V5.3 operating systems. The p5-570 servers are configured with multiple logical partitions and provide the client with increased server processing power compared to the other server environment. The p5-570 server platform runs the client's enterprise management information systems (MIS), including its large SAP® environment and a number of enterprise planning modules.

3.4.3 Benefits

The IBM and Direct Systems Support solution enabled the client to consolidate its MIS workloads from a large environment from an IBM competitor to a smaller, more powerful IBM System p5 server platform. Compared to the previous environment, the p5-570 servers offer lower maintenance costs, decreased power and facilities expenses, and a smaller number of processors, resulting in a significant reduction of software-licensing costs. Overall, the client estimates that it will reduce the TCO for its server environment by 20%.

The partition feature of the System p5 servers provides the client with an ease of management not possible with the previous environment. By leveraging the partition technology, the client can run different levels of the AIX 5L operating system on a single server. This capability enables it to run multiple applications on the same box without worrying about compatibility issues between the AIX 5L V5.2 and V5.3 operating systems.

3.4.4 Architectural design and diagrams

Here we describe the main points, followed by the diagrams showing the configuration:

- IBM has architected the sizing, including 10% additional resources for year-end processing, plus 50% growth for production and two concurrent environments.
- The total number of IBM p5-570 16-core servers is nine.

- The DR server is (1) IBM System p5 550 4-core with 16 GB memory.
- This architecture assures that the various transaction systems are capable of growing 50% and executing at peak workload.
- IBM has ensured that there is flexibility for production environments to be allocated resources from non-production environments on the same server.
- Each p5-570 server is delivered with 128 GB of memory, some with portions of memory delivered as Memory on Demand.
- With AIX 5L 5.3 and multithreading improved performance measured at 30%, IBM incorporated 20% improvement for the configuration.

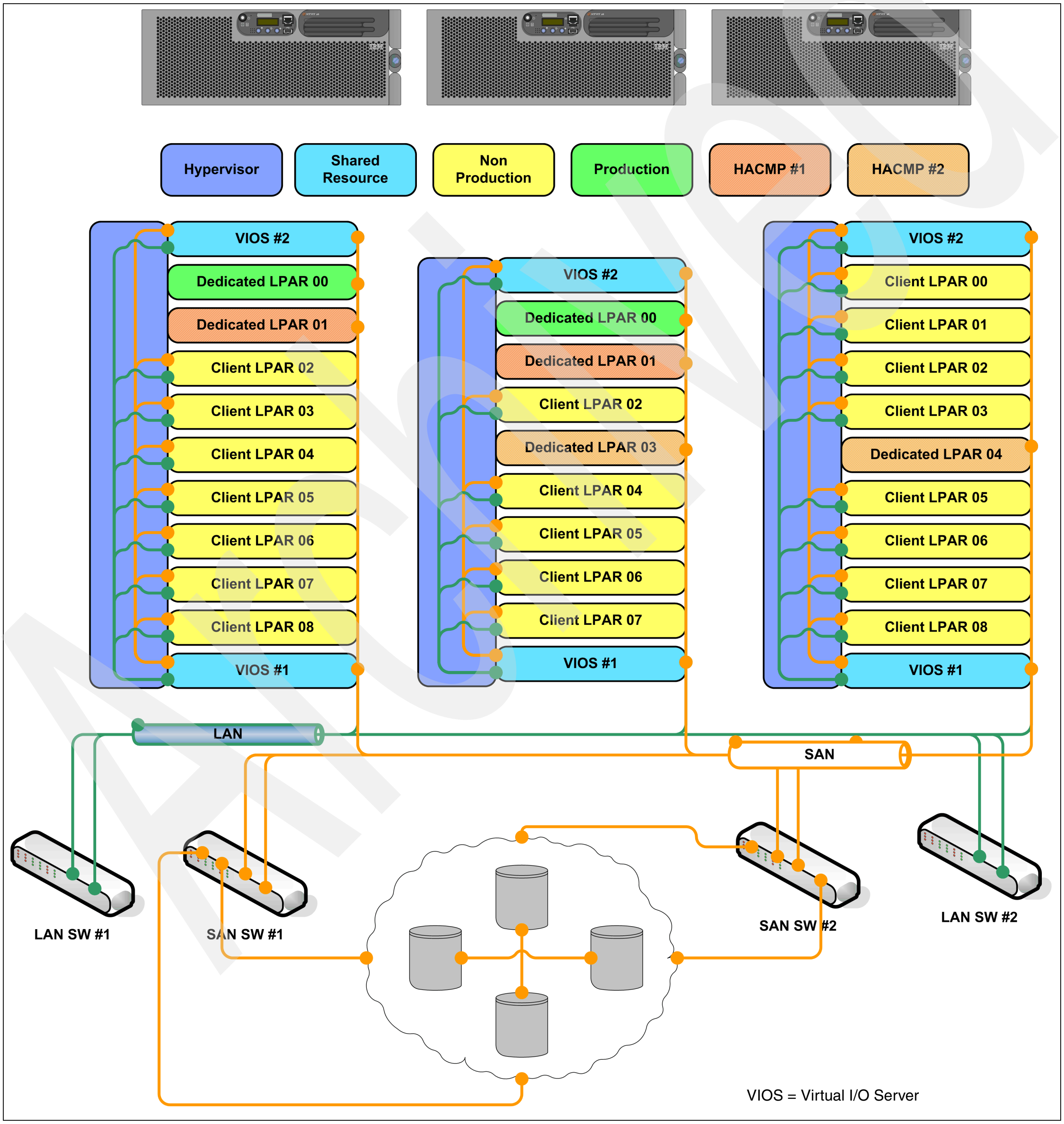
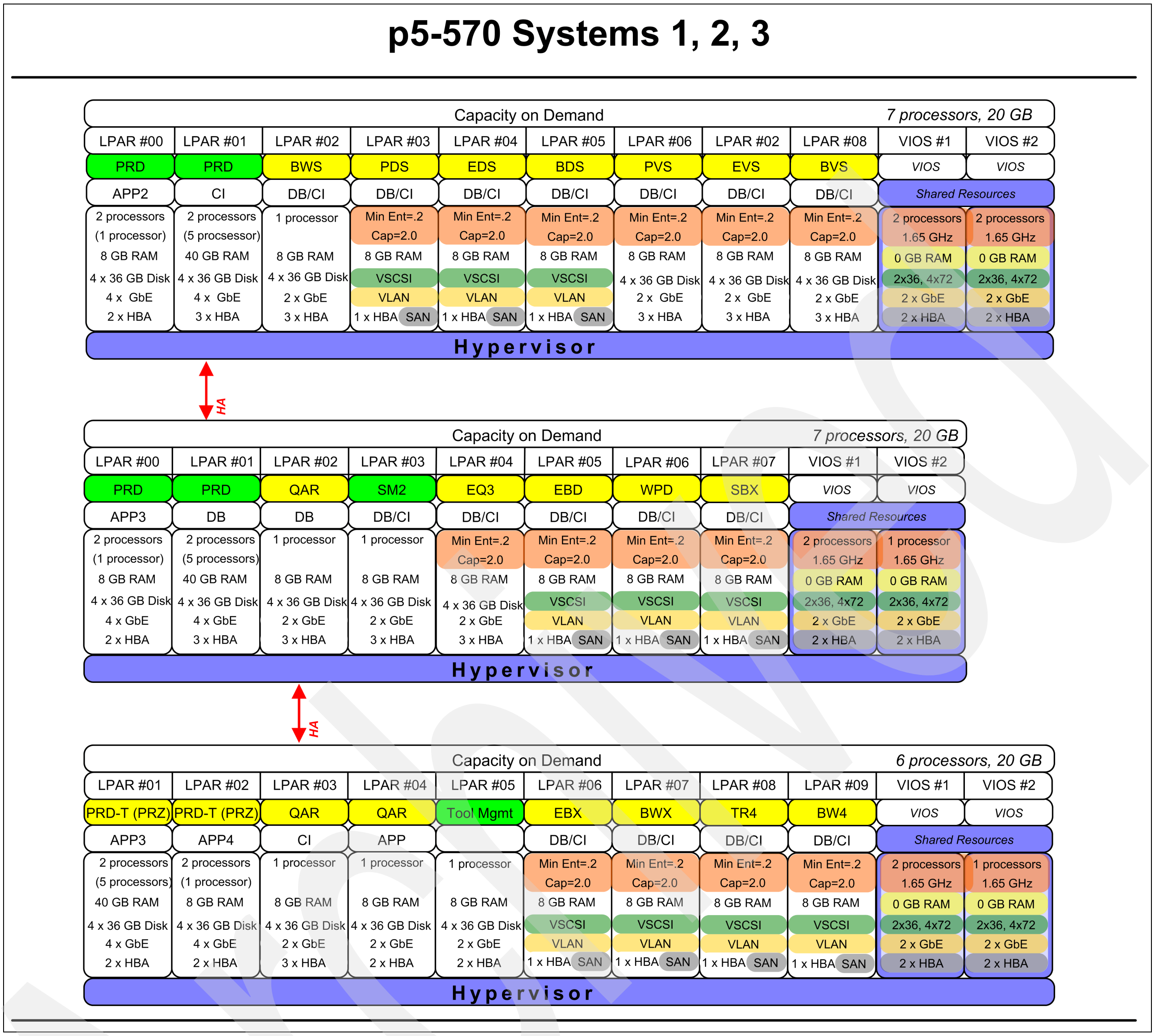


Figure 3-9   p5-570 complex with shared, dedicated resources and HACMP for serviceability



p5-570 Systems 1, 2, 3

Capacity on Demand									7 processors, 20 GB	
LPAR #00	LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05	LPAR #06	LPAR #02	LPAR #08	VIOS #1	VIOS #2
PRD	PRD	BWS	PDS	EDS	BDS	PVS	EVS	BVS	VIOS	VIOS
APP2	CI	DB/CI	DB/CI	DB/CI	DB/CI	DB/CI	DB/CI	DB/CI	Shared Resources	
2 processors	2 processors	1 processor	Min Ent=.2	Min Ent=.2	Min Ent=.2	Min Ent=.2	Min Ent=.2	Min Ent=.2	2 processors	2 processors
(1 processor)	(5 procsessor)		Cap=2.0	Cap=2.0	Cap=2.0	Cap=2.0	Cap=2.0	Cap=2.0	1.65 GHz	1.65 GHz
8 GB RAM	40 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	0 GB RAM	0 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	VSCSI	VSCSI	VSCSI	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	2x36, 4x72	2x36, 4x72
4 x GbE	4 x GbE	2 x GbE	VLAN	VLAN	VLAN	2 x GbE	2 x GbE	2 x GbE	2 x GbE	2 x GbE
2 x HBA	3 x HBA	3 x HBA	1 x HBA SAN	1 x HBA SAN	1 x HBA SAN	3 x HBA	3 x HBA	3 x HBA	2 x HBA	2 x HBA
Hypervisor										

Capacity on Demand								7 processors, 20 GB	
LPAR #00	LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05	LPAR #06	LPAR #07	VIOS #1	VIOS #2
PRD	PRD	QAR	SM2	EQ3	EBD	WPD	SBX	VIOS	VIOS
APP3	DB	DB	DB/CI	DB/CI	DB/CI	DB/CI	DB/CI	Shared Resources	
2 processors	2 processors	1 processor	1 processor	Min Ent=.2	Min Ent=.2	Min Ent=.2	Min Ent=.2	2 processors	1 processor
(1 processor)	(5 processors)			Cap=2.0	Cap=2.0	Cap=2.0	Cap=2.0	1.65 GHz	1.65 GHz
8 GB RAM	40 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	0 GB RAM	0 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	VSCSI	VSCSI	VSCSI	2x36, 4x72	2x36, 4x72
4 x GbE	4 x GbE	2 x GbE	2 x GbE	2 x GbE	VLAN	VLAN	VLAN	2 x GbE	2 x GbE
2 x HBA	3 x HBA	3 x HBA	3 x HBA	3 x HBA	1 x HBA SAN	1 x HBA SAN	1 x HBA SAN	2 x HBA	2 x HBA
Hypervisor									

Capacity on Demand									6 processors, 20 GB	
LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05	LPAR #06	LPAR #07	LPAR #08	LPAR #09	VIOS #1	VIOS #2
PRD-T (PRZ)	PRD-T (PRZ)	QAR	QAR	Tool Mgmt	EBX	BWX	TR4	BW4	VIOS	VIOS
APP3	APP4	CI	APP		DB/CI	DB/CI	DB/CI	DB/CI	Shared Resources	
2 processors	2 processors	1 processor	1 processor	1 processor	Min Ent=.2	Min Ent=.2	Min Ent=.2	Min Ent=.2	2 processors	1 processors
(5 processors)	(1 processor)				Cap=2.0	Cap=2.0	Cap=2.0	Cap=2.0	1.65 GHz	1.65 GHz
40 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	0 GB RAM	0 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	VSCSI	VSCSI	VSCSI	VSCSI	2x36, 4x72	2x36, 4x72
4 x GbE	4 x GbE	2 x GbE	2 x GbE	2 x GbE	VLAN	VLAN	VLAN	VLAN	2 x GbE	2 x GbE
2 x HBA	2 x HBA	3 x HBA	2 x HBA	2 x HBA	1 x HBA SAN	1 x HBA SAN	1 x HBA SAN	1 x HBA SAN	2 x HBA	2 x HBA
Hypervisor										

Figure 3-10 Hardware breakdown of p5-570 servers

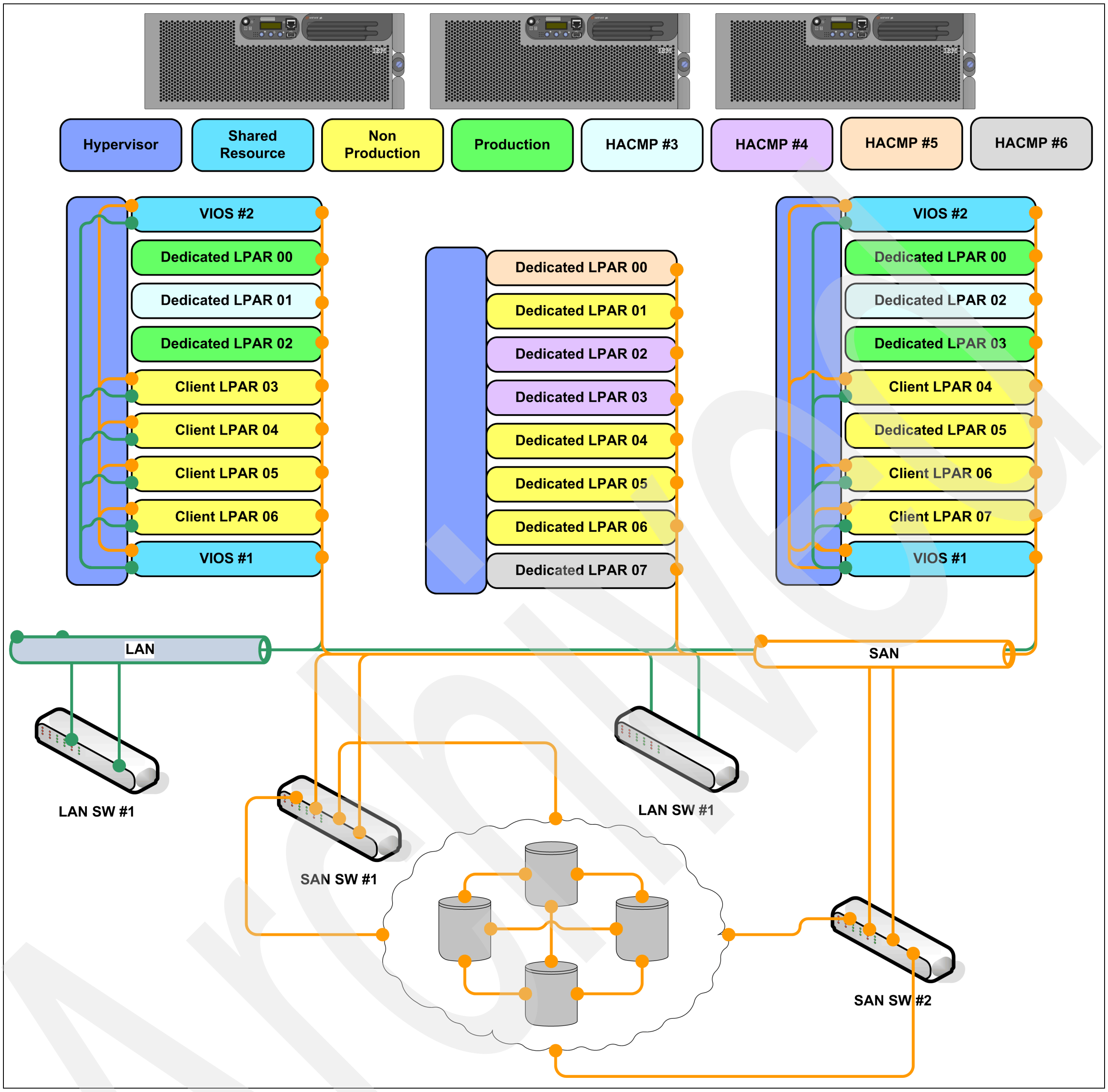
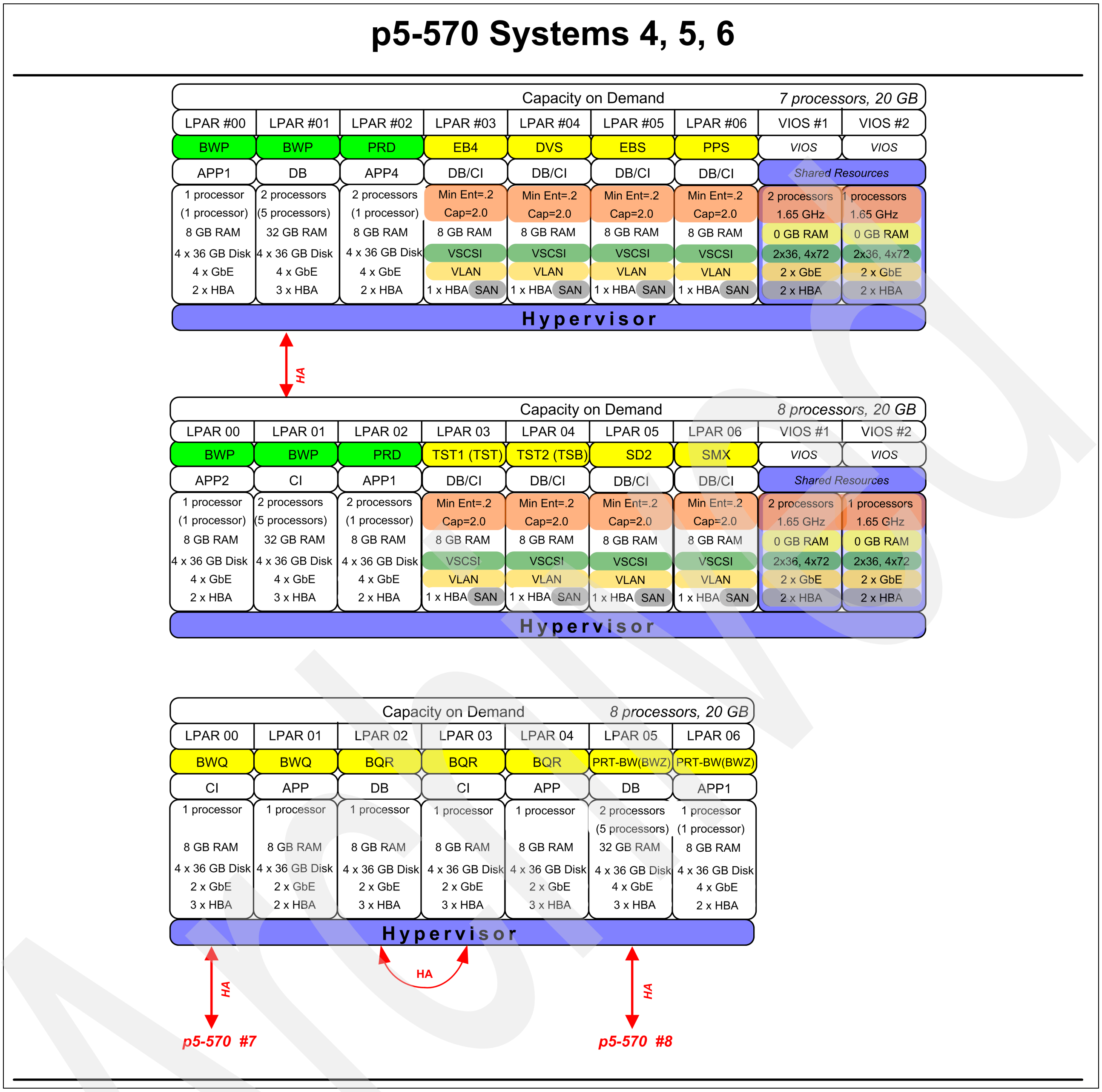


Figure 3-11 p5-570 servers with dedicated and virtual resources using HACMP



p5-570 Systems 4, 5, 6

Capacity on Demand — *7 processors, 20 GB*

LPAR #00	LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05	LPAR #06	VIOS #1	VIOS #2
BWP	BWP	PRD	EB4	DVS	EBS	PPS	*VIOS*	*VIOS*
APP1	DB	APP4	DB/CI	DB/CI	DB/CI	DB/CI	*Shared Resources*	
1 processor (1 processor) 8 GB RAM 4 x 36 GB Disk 4 x GbE 2 x HBA	2 processors (5 processors) 32 GB RAM 4 x 36 GB Disk 4 x GbE 3 x HBA	2 processors (1 processor) 8 GB RAM 4 x 36 GB Disk 4 x GbE 2 x HBA	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	2 processors 1.65 GHz 0 GB RAM 2x36, 4x72 2 x GbE 2 x HBA	1 processors 1.65 GHz 0 GB RAM 2x36, 4x72 2 x GbE 2 x HBA

Capacity on Demand — *8 processors, 20 GB*

LPAR 00	LPAR 01	LPAR 02	LPAR 03	LPAR 04	LPAR 05	LPAR 06	VIOS #1	VIOS #2
BWP	BWP	PRD	TST1 (TST)	TST2 (TSB)	SD2	SMX	*VIOS*	*VIOS*
APP2	CI	APP1	DB/CI	DB/CI	DB/CI	DB/CI	*Shared Resources*	
1 processor (1 processor) 8 GB RAM 4 x 36 GB Disk 4 x GbE 2 x HBA	2 processors (5 processors) 32 GB RAM 4 x 36 GB Disk 4 x GbE 3 x HBA	2 processors (1 processor) 8 GB RAM 4 x 36 GB Disk 4 x GbE 2 x HBA	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	Min Ent=.2 Cap=2.0 8 GB RAM VSCSI VLAN 1 x HBA SAN	2 processors 1.65 GHz 0 GB RAM 2x36, 4x72 2 x GbE 2 x HBA	1 processors 1.65 GHz 0 GB RAM 2x36, 4x72 2 x GbE 2 x HBA

Capacity on Demand — *8 processors, 20 GB*

LPAR 00	LPAR 01	LPAR 02	LPAR 03	LPAR 04	LPAR 05	LPAR 06
BWQ	BWQ	BQR	BQR	BQR	PRT-BW(BWZ)	PRT-BW(BWZ)
CI	APP	DB	CI	APP	DB	APP1
1 processor 8 GB RAM 4 x 36 GB Disk 2 x GbE 3 x HBA	1 processor 8 GB RAM 4 x 36 GB Disk 2 x GbE 2 x HBA	1 processor 8 GB RAM 4 x 36 GB Disk 2 x GbE 3 x HBA	1 processor 8 GB RAM 4 x 36 GB Disk 2 x GbE 3 x HBA	1 processor 8 GB RAM 4 x 36 GB Disk 2 x GbE 3 x HBA	2 processors (5 processors) 32 GB RAM 4 x 36 GB Disk 4 x GbE 3 x HBA	1 processor (1 processor) 8 GB RAM 4 x 36 GB Disk 4 x GbE 2 x HBA

Figure 3-12 Hardware breakdown of p5-570 servers

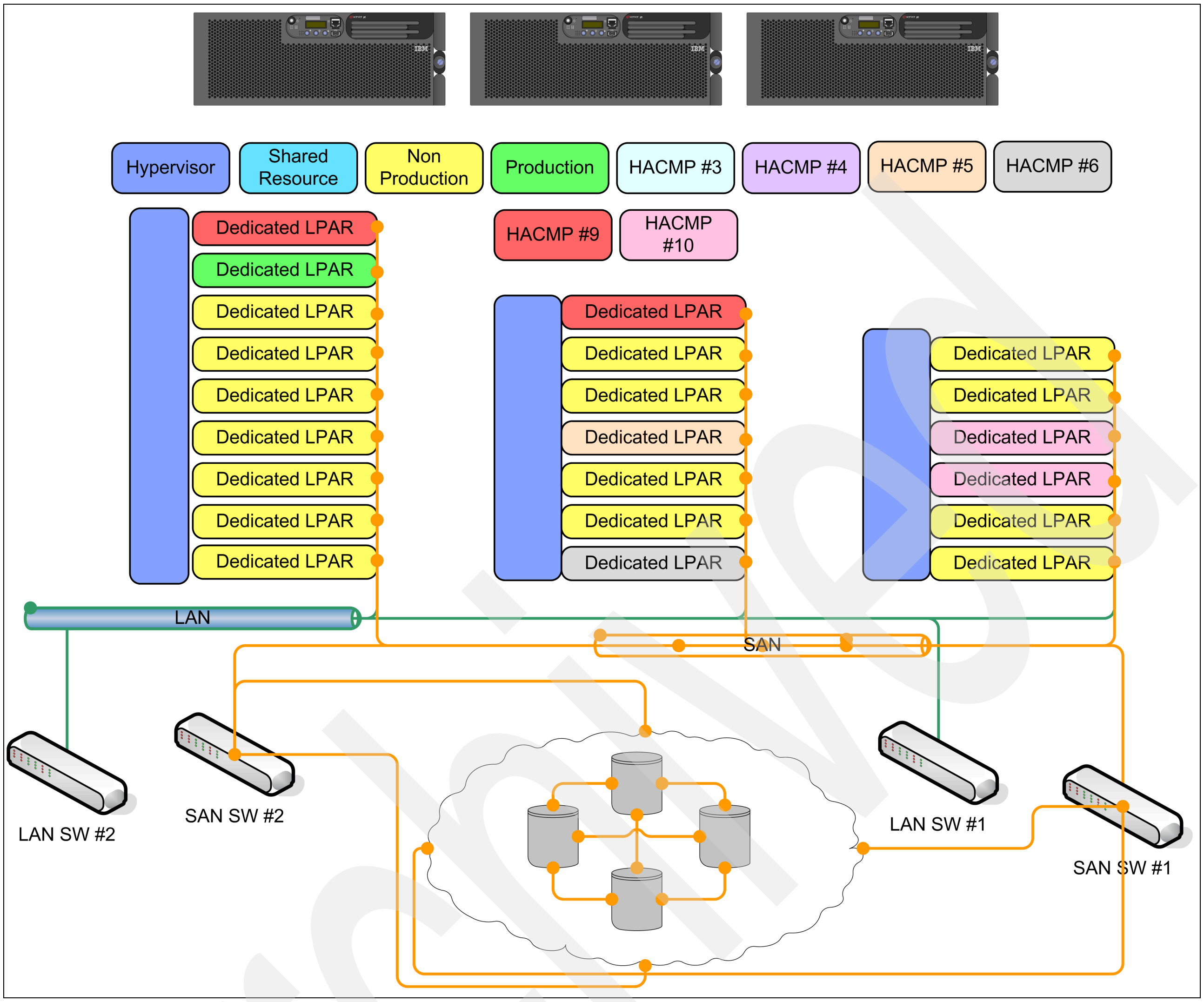
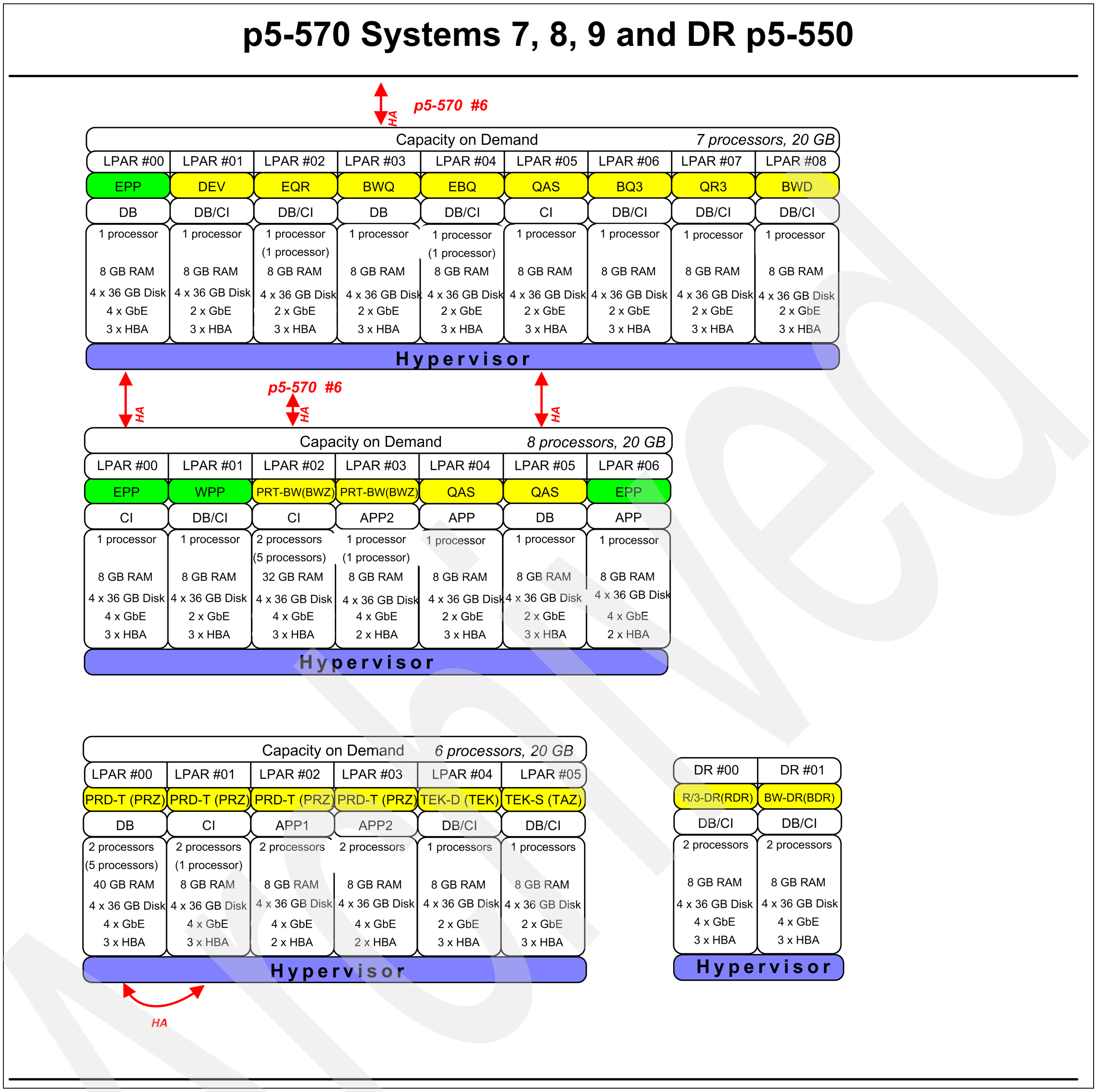


Figure 3-13 p5-570 servers with dedicated and virtual resources using HACMP and all without Virtual I/O Servers



p5-570 Systems 7, 8, 9 and DR p5-550

Capacity on Demand							7 processors, 20 GB	
LPAR #00	LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05	LPAR #06	LPAR #07	LPAR #08
EPP	DEV	EQR	BWQ	EBQ	QAS	BQ3	QR3	BWD
DB	DB/CI	DB/CI	DB	DB/CI	CI	DB/CI	DB/CI	DB/CI
1 processor	1 processor	1 processor (1 processor)	1 processor	1 processor (1 processor)	1 processor	1 processor	1 processor	1 processor
8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk
4 x GbE	2 x GbE	2 x GbE	2 x GbE	2 x GbE	2 x GbE	2 x GbE	2 x GbE	2 x GbE
3 x HBA	3 x HBA	3 x HBA	3 x HBA	3 x HBA	3 x HBA	3 x HBA	3 x HBA	3 x HBA

Capacity on Demand					8 processors, 20 GB	
LPAR #00	LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05	LPAR #06
EPP	WPP	PRT-BW(BWZ)	PRT-BW(BWZ)	QAS	QAS	EPP
CI	DB/CI	CI	APP2	APP	DB	APP
1 processor	1 processor	2 processors (5 processors)	1 processor (1 processor)	1 processor	1 processor	1 processor
8 GB RAM	8 GB RAM	32 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk
4 x GbE	2 x GbE	4 x GbE	4 x GbE	2 x GbE	2 x GbE	4 x GbE
3 x HBA	3 x HBA	3 x HBA	2 x HBA	3 x HBA	3 x HBA	2 x HBA

Capacity on Demand				6 processors, 20 GB	
LPAR #00	LPAR #01	LPAR #02	LPAR #03	LPAR #04	LPAR #05
PRD-T (PRZ)	PRD-T (PRZ)	PRD-T (PRZ)	PRD-T (PRZ)	TEK-D (TEK)	TEK-S (TAZ)
DB	CI	APP1	APP2	DB/CI	DB/CI
2 processors (5 processors)	2 processors (1 processor)	2 processors	2 processors	1 processors	1 processors
40 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM	8 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk	4 x 36 GB Disk
4 x GbE	4 x GbE	4 x GbE	4 x GbE	2 x GbE	2 x GbE
3 x HBA	3 x HBA	2 x HBA	2 x HBA	3 x HBA	3 x HBA

DR #00	DR #01
R/3-DR(RDR)	BW-DR(BDR)
DB/CI	DB/CI
2 processors	2 processors
8 GB RAM	8 GB RAM
4 x 36 GB Disk	4 x 36 GB Disk
4 x GbE	4 x GbE
3 x HBA	3 x HBA

Figure 3-14 Hardware breakdown of p5-570 servers

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this Redpaper.

IBM Redbooks

For information about ordering these publications, see "How to get IBM Redbooks" on page 84. Note that some of the documents referenced here may be available in softcopy only.

- *Advanced POWER Virtualization on IBM eServer p5 Servers: Architecture and Performance Considerations*, SG24-5768
- *Advanced POWER Virtualization on IBM System p5*, SG24-7940
- *Effective System Management Using the IBM Hardware Management Console for pSeries*, SG24-7038
- *IBM System p Advanced POWER Virtualization Best Practices*, REDP-4194
- *IBM System p5 Approaches to 24x7 Availability Including AIX 5L*, SG24-7196
- *IBM TotalStorage DS300 and DS400 Best Practices Guide*, SG24-7121
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061
- *NIM: From A to Z in AIX 4.3*, SG24-5524
- *Partitioning Implementations for IBM eServer p5 Servers*, SG24-7039
- *The IBM TotalStorage DS6000 Series: Concepts and Architecture*, SG24-6471
- *The IBM TotalStorage DS8000 Series: Concepts and Architecture*, SG24-6452

Online resources

These Web sites are also relevant as further information sources:

- Detailed documentation about the Advanced POWER Virtualization feature and the Virtual I/O Server

  `http://techsupport.services.ibm.com/server/vios/documentation/home.html`

- Latest *Multipath Subsystem Device Driver User's Guide*

  `http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&uid=ssg1S7000303`

- IBM System Planning Tool

  `http://www.ibm.com/servers/eserver/support/tools/systemplanningtool/`

- Virtual I/O Server home page

  `http://techsupport.services.ibm.com/server/vios/home.html`

- Capacity on Demand

  `http://www.ibm.com/systems/p/cod/`

- Subsystem Device Driver Path Control Module (SDDPCM) software download page

  http://www.ibm.com/support/docview.wss?uid=ssg1S4000201
- Subsystem Device Driver (SDD) software download page

  http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&dc=D430&uid=ssg1S4000065&loc=en_US&cs=utf-8&lang=en
- HACMP for System p

  http://www.ibm.com/systems/p/software/hacmp.html
- Virtual I/O Server supported hardware

  http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/VIOS_datasheet_063006.html
- IBM Systems Workload Estimator

  http://www-304.ibm.com/jct01004c/systems/support/tools/estimator/index.html
- IBM Systems Hardware Information Center

  http://publib.boulder.ibm.com/infocenter/eserver/v1r3s/index.jsp
- IBM eServer pSeries and AIX Information Center

  http://publib16.boulder.ibm.com/pseries/index.htm
- Virtualization wiki

  http://www.ibm.com/collaboration/wiki/display/virtualization/Home
- IBM Advanced POWER Virtualization on IBM System p Web page

  http://www.ibm.com/systems/p/apv/

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM®

Advanced POWER Virtualization on IBM System p Virtual I/O Server Deployment Examples

Redpaper

Scenarios for quick deployment of virtualization on IBM System p

Real world virtualization examples

A reference for system architects

The *Advanced POWER Virtualization on IBM System p Virtual I/O Server Deployment Examples*, REDP-4224, publication provides a number of high-level system architecture designs using the Advanced POWER Virtualization feature available on IBM System p5 servers. These high-level architecture designs are referred to as scenarios and they show different configurations of the Virtual I/O Server and client partitions to meet the needs of various solutions.

The Advanced POWER Virtualization feature is very flexible and can support several configurations designed to provide cost savings and improved infrastructure agility. We selected the scenarios described in this paper to provide some specific examples to help you decide on your particular implementation and perhaps extend and combine the scenarios to meet additional requirements.

This publication is targeted at architects who are interested in leveraging IBM System p virtualization using the Virtual I/O Server to improve your IT solutions. System architects can use the scenarios as a basis for developing their unique virtualization deployments. Business Partners can review these scenarios to help them understand the robustness of this virtualization technology and how to integrate it with their solutions. Clients can use the scenarios and examples to help them plan improvements to their IT operations.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks